Journal of Advanced Research in Image Processing and Applications
Volume 4, Issue 1 - 2021, Pg. No. 7-11
Peer Reviewed Journal

Article

A Methodical Analysis on Image Captioning Techniques

Pilli Vijay Lakshmanrao[1], Kotta Vijay Kumar[2], Mekathoti Tejaswi[3], K Lakshman Rao[4]

[1,2]B.Tech 5th Semester Students, Dept. of CSE, GMR Inst. of Tech., Rajam, Andhra Pradesh, india.
[3]Associate Professor, Dept. of CSE, GMR Inst. of Tech., Rajam, Andhra Pradesh, india.

I N F O

Corresponding Author:
Pilli. Vijay Lakshmanrao, Dept. of CSE, GMR Inst. of Tech., Rajam, Andhra Pradesh, india.
E-mail Id:
vijaylakshmanarao@gmail.com
Orcid Id:
https://orcid.org/0000-0003-1892-6023
How to cite this article:
Lakshmanrao PV, Kumar KV, Tejaswi M et al. Methodical Analysis on Image Captioning Techniques. *J Adv Res Image Proc Appl* 2021; 4(1): 7-11.

Date of Submission: 2021-04-20
Date of Acceptance: 2021-05-10

A B S T R A C T

Image captioning, it is the process of describing an image given by connecting techniques like Computer Vision (CV) and Natural Language Processing (NLP). Image captioning is done by three different methods Object Detection technique, Encoder Decoder technique and Attention Mechanism. Each of these techniques have different approaches to obtain image captioning.

This paper handles the approaches and methodologies that are followed by the image.

Natural Language Processing (NLP) linked with Computer Vision (CV) is the bridge for Image Captioning i.e., human-machine interaction (Neural network techniques as CNN and RNN), Artificial Intelligence (AI) as first of its kind. Improving the effectiveness of image using image attributes based on semantic attention, considering the attributes that contains the high-level awareness of image content and particular schematics of correlating captioning words. It is big task to convert visible data into text manner, on the either side of the coin image captioning algorithm is needed to amend the rough schematic concept to human like natural language descriptions step by step. Multi-level features fusion might be a better solution for image captioning had attracted numerous research interests and huge number of models are being proposed. The substantial advances in the deep neural networks attention model with spatial region is on focus.Neural Networks and computer vision approaches for image captioning in different models are presented.

Keywords: Image Caption, Neural Networks, Encoder-Decoder, Long Short Term Memory, Description

Introduction

Deriving a meaningful sentence for the given image is called image captioning. It is very fast-growing technology in recent days. Day to day new methods with various frameworks are introduced to achieve adequate results. Both Syntactic and semantic understanding are essential for sentence generation. This image captioning can help the visually impaired (vision problem), for illiterates, fields of security and military and for describing medical images.

The three major approaches to capture and describe the image which give better results in caption generation are Object Detection, Encoder Decoder mechanism under the concept of CNN and LSTM and next approach is Attention mechanism was a extension of Encoder Decoder .

The object detection which follows different applications to generate caption to image. The given image is divided

Copyright (c) 2021 Advanced Research Publications

into multiple smaller pixels. The relation between each pixel is detected and a meaningful sentence is generated.

The Encoder-Decoder approach uses the architecture of encoder with decoder which involves CNN and as well as RNN. CNN used to find the extraction of features of input image and RNN is used to helps to find the generation of sentences for a given input image. The Encoder used for sentence generation. The encoder used to map input to the real sentences.

As a problem has many strategies, Image captioning Technology also has three different methods.

The thoroughly researched and processed methods to attain image captioning is:

1. Encoder-Decoder
2. Object Detection
3. Attention Mechanism

Literature Survey

Xinyu Xiao, Ling feng Wang explained that the layers involved in encoder-decoder model.and these layers hepls in such a way that how much it effectively fuses visual semantics and as well textual semantics for generating the appropriate sentence or captions for the input images.

Harshit Parikh, Harsh Sawant, Bhautik Parmar, Rahul Shah, Santosh Chapaneri, Deepak Jaiswal explained the Gated Recurrent Unit algorithm for generating the captions for an input image.GRU was very helpful to train the project in very less time since it has less number of gate.Though it require less time to train but it gives relevant descriptions for an input image.

Imane Allaaouzi, M. Ahmed, B. Benamrou intensed to develop the image captioning int hte field of medical domain. they provided an overview of existing works, datasets and evaluation metrics to develop the image captioning properly in the medical field.

Kang Tong and Yiquaqn Wu researched to detect smaller instances of object in the given input image. They used tasks such as computer vision, object tracking, instance segmentation, image captioning, action recognition and scene understanding.

Xin hang song, Cheng Peng chen tried to solve the intraclass diversity and interclass community. They used two types of representation COOR and SOOR.

Payal Mittal, Aakash deep Sharma, Raman Singh worked on aerial datasets that consists of images took from longer distances I.e., from drones or planes. They found the critical issues in object detection on drone platform like small object detection, occlusion, large scale variation, class imbalance.

Concept behind Image Captioning

Convolution Neural Network (CNN)

CNN usually to classify the input image into different categories (E.g., Dog, Cat, Tiger, Lion) with the help of convolution layers. As the image passed into filter layer of convolution neural network, it was the responsible for classifying the input image accordingly. Edge detecting is one of the actions done by each layer. With the help of image resolution, the input image is observed as array of pixels by computer. An input RGB image is represents with 6 x 6 x 3 array of matrix (3 refers to RGB values) and an input image grayscale is represents with 4 x 4 x 1 array of matrix. To train and test involves deep learning CNN models for the input image will be passed into a convolution layers which is having series of filters and then pass it through Pooling, followed by fully connected layers (FC). This CNN helps to classify the input image to object with values that are probabilistic between 0 and 1 by applying Soft Max function.

Long Short-Term Memory

RNN failed in train the large number of inputs or long-term sequences. It leads to vanishing gradient problem (weights are negligible) and exploding gradient problem (weights are big) which might very difficult that gradients propagating from back through large number of recurrent layers. LSTM has special features like it can easily forget previous step when new information is given to LSTM, it can easily update hidden state. In sequence generation, LSTM plays a important role to generate the output in sequences and as well machine translation uses the concept of LSTM.

LSTM hidden layer consists of one cell and three gates.

1. The cell is used to store data is ‘Memory cell’.
2. LSTM is used to help forget previous step through ‘Forget Gate’.
3. LSTM is used to take input through ‘Input Gate’.
4. LSTM helps to give output through ‘Output Gate’.

Image captioning using Deep Hierarchical Encoder-Decoder Network

Four modules are involved in this hierarchical network .In which, there are two encoders and two decoders which are mainly used to extract the features of the images and describing the images.

1. Encoder of Deep CNN.
2. Encoder of Sentence-LSTM (S-LSTM).
3. Vision-Sentence Embedding LSTM decoder(VSE-LSTM).
4. Semantic Fusion LSTM decoder(SF-LSTM).

Deep CNN Encoder

When the input image is given to encoder decoder model

as it was a first module in this model, this deep CNN used to encode the features of the image. Actually, this Deep CNN helps to find the particular objects to which region related and to find edge detection to know the difference of one input image to another input image.

Sentence-LSTM Encoder(S-LSTM)

The S-LSTM encoder is second module in encoder decoder. Order of one-hot vectors extracted from the input image passed to the S-LSTM. This module updates the hidden state, this later involves encoding to sentence from the one hot vector representation of image.

Vision-Sentence Embedding LSTM Decoder (VSE-LSTM):

Vision Sentence Embedding LSTM is the third module in encoder decoder model, in which it allows for fusing the CNN features of visuals from the module of Deep CNN and SLSTM feature sentences from the module of S-LSTM and transforming to a joint semantic space.

Semantic Fusion LSTM Decoder (SF-LSTM)

A Semantic Fusion LSTM Decoder (SF-LSTM) is the fourth module of encoder-decoder. This module helps to get the target sentences which are appropriate captions for the input image.

Image Captioning using Attention Mechanism

Global Attention Mechanism Approach

In this attention mechanism approach takes consideration of all encoder hidden states to drive the Context Vector (c(t)). To calculate the context vector (c(t)) , we need to compute Alignment Vector (a(t)) of variable lengths. The derivation of alignment vector is done by computing the similarities of source and target hidden states that are through the score function we can calculate the similarities of the states of the image.

Deriving the Hidden Encoded States

The hidden state is produced by the encoder for every element in the sequence of input.

RNN Decoder

The new hidden state is generated by the RNN Decoder by taking hidden state and previous decoder output as the input for that time stamp.

Alignment Scores Calculation

The alignment scores are calculated by the newly generated encoder and decoder hidden states.

Soft maxing the Alignment Scores

Hidden states of encoder alignment scores were combined and it was represented with vector and then soft maxed technique is used.

Context Vector Calculation

Context vector is formed by the combination of the alignment scores and encoded hidden states.

Production of the Final Output

The decoder hidden state is combined with the context vector. This combination as passed through a fully connected layer produces a new output.

Local Attention Mechanism Approach

This approach is mainly focusing on words that are involved in forming a sentence. Whenever this approach is used for image captioning model it results to cost expenditure is more and practically it will not give better results. In order to overcome this deficiency local attention model is used. Small subset of the encoder hidden states focused mainly in this approach. Attention is mainly placed only on a few source positions.

Deriving the Hidden Encoded States

The hidden state is produced by the encoder for every element in the sequence of input.

Calculating Alignment Scores

Alignment scores are calculated between two states.so the scores are calculated between hidden state of previous decoder and hidden state of encoder.

Soft maxing the Alignment Scores

Hidden states of encoder alignment scores were combined and it was represented with vector and then soft maxed technique is used.

Context Vector Calculation

Context vector is formed by the combination of the alignment scores and encoded hidden states.

Production of the Final Output

The decoder hidden state is combined with the context vector. This combination as passed through a fully connected layer produces a new output.

Image Captioning using object Detection

The main theme of the object detecting mechanism is to detecting objects and establish relationship between them. It combines both Computer Vision and Image processing. The task that involves computer vision and Natural language processing to describe picture in human understandable language is called Image caption generator system.

Object Detection takes place in two steps:

Image is divided into various smaller pieces.

Processing the pieces in to an image classifier to check the given image has objects or not.

For dividing the Image in to parts we have algorithms like:

1. Sliding window Algorithm
2. Image segmentation
3. Selective search Algorithm

Sliding Window Algorithm

1. In this Algorithm the given input image is converted into Smaller pieces generating grids or windows.
2. After dividing the image into parts, we select the grid or window of specific size.
3. We process the selected grid cell along the image and convert the piece of the image in the gird cell for predicting the output.
4. Next, we glide the grid cell along slide two and then convert the following part of the image.
5. This is the process which we convert entire image and repeating the same process with distinct grid cells.

Image Segmentation

In this algorithm the image is divided based on pixels. Similar pixel images are:

1. Put together to form a region. Next the regions formed are given to CNN classifier to recognize the image It involves dividing the image into segments which represents objects and set of pixels. There are three levels for image analysis:
2. Object detection method of detecting objects for given image and representing them in rectangular shape e.g., a human or a goat.
3. Classification classifying the given whole image into classes such as “people”, “animals”, “outdoors”
4. Segmentation recognizing the pieces of the image and concluding to which object they belong to. The basic knowledge needed for performing object detection is Segmentation.

Selective Search Algorithm

The most important thing in the Object detection is the object localization. The issue of object localization is the most crucial thing of object detection. Sliding window is one of the process which we follow in which windows of different size are used to locate objects for the given image. The method is called Exhaustive search. This method is so expensive because the given image needs to be searched for object in thousands of windows, for small objects too. Some changes are done like, window sizes are noted in different ratios (behalf of changing the pixels). But due to multiple number of windows it is not so effective. Both Exhaustive search and Segmentation (process to differentiate objects based on different shapes in the image by giving them different colors) are used by Selective Search algorithm.

Conclusion

In this growing world of technology image captioning is one of the highly centred technology and developing to a great extent to meet all the real-world requirements of the image captioning is used in different ways, for visually impaired people i.e., actually impaired person cannot see anything in front of them, by using the image captioning techniques the captions that get by image captioning was read to the blind person by using natural language processing ,reading anything like newspaper for illiterate people which was written in common natural understanding of the language ,self-driving cars usually captures the image of scenario and act according to that scenario and automatic image caption can make google image search as efficient as google search. Image captioning can be used to capture vehicles that are not following traffic rules. Image captioning plays a crucial role in the future generation. Many scientists are used to work on this experiment for a better knowledge and better implementation of the image captioning.

References

1. Xiao X, Wang LF, Ding K et al. Deep Hierarchical Encoder-Decoder Network for Image Captioning. *IEEE Transactions on Multimedia* 2019.
2. Parikh H, Sawant H, Parmar B et al. Encoder-Decoder Architecture for Image Caption Generation. *IEEE 3rd International Conference on Communication System*, Computing and IT Applications, 2020.
3. Allaaouzi I, Ahmed MB, Benamrou B et al. Automatic Caption Generation for Medical Images’ Proceedings of the 3rd international conference on smart city applicatiuons 2018.
4. Wang B, Wang C, Zhang Q et al. Cross-Lingual Image Caption Generation Based on Visual Attention Model. *IEEE Access* 2020; 8.
5. Cheng L, Wei W, Mao X et al. Stack-VS: Stacked Visual-Semantic Attention for Image Caption Generation. in IEEE Access 2018; 8.
6. Zhang Z, Wu Q, Wang Y et al. High-Quality Image Captioning with Fine-Grained and Semantic-Guided Visual Attention. in *IEEE Transactions on Multimedia* 2018; 21: 1681-1693.
7. Huang Y, Chen J, Ouyang W et al. Image Captioning with End-to-End Attribute Detection and Subsequent Attributes Prediction. *IEEE Transactions on Image Processing* 2020; 29.
8. Wu C, Yuan S, Cao H et al. Hierarchical Attention-Based Fusion for Image Caption with Multi-Grained Rewards. *IEEE Access* 2000; 8.
9. Song XH, Chen CP, Chen GW. Image captioning with object-to-object relations. *IEEE Transactions on Image Processing* 2019; 29.
10. Huang Y, Ouyang W, Xue Y. Image captioning with end-to-end attribute detection. *IEEE Transactions on Image processing* 2020; 29.

11. Tong K, wu Y, Zhou F. Recent Advances in small Object detection based on deep Learning. Science Direct, *Image and Vision Computing* 2020; 97.
12. Mittal P, Sharma AD, Singh R. Deep Learning based object detection in low altitude UAV datasets", Science Direct, Image and vision Computing, 2020.
13. Hoxha G, Melgani F, Salaghenauffi J. A New CNN-RNN frame work for remote sensing Image captioning. *IEEE, Communication and image captioning* 2020.