International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Bi-directional Long Short-Term Memory with Convolutional Neural Network Approach for Image Captioning

Suramya Patel‡* and Shilpa Gite‡

‡CS/IT Department, Symbiosis Institute of Technology, SIU, Pune, India

Received 01 Sept 2017, Accepted 01 Nov 2017, Available online 10 Nov 2017, **Vol.7, No.6 (Nov/Dec 2017)**

Abstract

Picture Captioning is a testing assignment that has required a lot of information as highlight extraction to accomplish elite. In this paper, we introduce a novel neural systems design that Convert Image into Sentences utilizing a half and half bidirectional LSTM with CNN Approach, taking out the requirement for most element Engineering. We picture the development of bidirectional LSTM inside states after some time and subjectively investigate how our models make an interpretation of picture to sentence. Our proposed models are assessed on subtitle era and picture sentence recovery errands from Available Dataset. Convolutional neural networks (CNN) have turned out to be mainstream in picture handling for include extraction. We show that bidirectional LSTM with CNN Approach accomplish profoundly Performance and Significantly beat late strategies on Image Captioning.

Keywords: *Bi- Directional LSTM, CNN, Image Captioning, Computer Vision, Natural Language Processing, Context Awareness*

1. Introduction

In Image preparing, Image subtitling and producing sentence is testing and critical research region. Programmed and cost-sparing naming of the a huge number of pictures transferred to the Internet consistently that unrealistic without the era of inscriptions from pictures. There are two of the real fields in Artificial Intelligence field that unites cutting edge models. One Natural Language Processing and second Computer Vision. Utilizing bidirectional will run your contributions to two routes, one from past to future and one from future to past. What varies this approach from unidirectional is that in the BLSTM that runs in reverse you protect data from the future and utilizing the two shrouded states consolidated you can save data from both past and future [Cheng Wang, ACM, 2016]. Convolutional Neural Networks (CNN) have turned out to be Popular in picture preparing for highlight extraction (Feature extraction includes lessening the measure of assets required to portray a vast arrangement of information).

2. Literature Survey

Alex Graves and his group demonstrate that Bidirectional LSTM superior to both unidirectional LSTM and Recurrent Neural Networks (RNNs). In this paper they present Hidden Markov Model. Crossover BLSTM-HMM framework enhances a proportionate conventional HMM framework, and in addition unidirectional LSTM-HMM. Utilizing term weighted mistake, expanded its acknowledgment precision yet there are some constraint Using span weighted blunder, that diminished the order Performance of BLSTM [Alex graves, Neural Network,2005].

Alex Graves, Jurgen Schmidhuber are demonstrate that bidirectional systems outflank unidirectional ones, and Long Short Term Memory (LSTM) is substantially quicker and furthermore more exact than both standard Recurrent Neural Nets (RNNs) and time-windowed Multilayer Perceptrons (MLPs). however, Hidden variable which are use for determination process is not identified.so that variable is utilized for recognize a question [Alex Graves, arxiv, 2005].

Andrej Karpathy, Armand Joulin Introduce Structured max-edge target that can be utilized to enhance any protest discovery method. This display enables expectations for the picture to sentence recovery errand by utilizing fragmentation. Limitation of this paper is that Phrases portray some specific number of visual elements are not Modeled [Jeff Donahue, arxiv,2016,].

Andrej Karpathy, Li Fei-Fei proposed and produces characteristic dialect depictions of picture locales in light of feeble names from dataset of pictures and Sentences. however, this model can just create a

*Corresponding author **Suramya Patel** is a M.Tech Scholar; ORCID ID: 0000-0002-9686-1723 and **Shilpa Gite** is working as Assistant Professor

portrayal of one information cluster of pixels at a settled determination. Perform on high determination pixel required Large dataset and taking picture from dataset is tedious assignment [Andrej Karpathy, arxiv,2014]

Oriol Vinyals portray Generate Natural sentences depicting a picture by utilizing Image Captioning. In NLP Selection of important and the sifting through of immaterial data from jumbled visual scenes is troublesome. So this Model Require Visual consideration display [Oriol vinyals, IEEE,2015].

Kelvin Xu, Jimmy Lei Ba beat the restriction of past paper by present visual consideration demonstrate utilized standard backpropagation methods for going to redress protest. Utilizing this method perform both two way bearing operation. However, trouble to going to more than one protest. Distinguishing proof of question require visual consideration [kelvin xu, arxiv,2016].

Jonghwan Mun perform two operation.one picture guides visual in appropriate and second subtitles generator. So, manage picture in appropriate way and produce legitimate picture sentences utilizing direction catch generator. in any case, Sometimes Captures objects not portrayed in direction subtitle. Due to many question are recognized so its hard to oversee them [JongHwan Mun ,arxiv,2016].

Qing Sun present Bidirectional Beam Search (BIBS) a novel Fill-in-the-Blank Image Captioning assignment which utilized both past and future sentence structure to recreate sensible picture depictions. Yet, there are some impediment is that Time Consuming To choose appropriate Description in fill the blanks. To distinguish legitimate Path to fill the clear is troublesome [Qing sun, IEEE,2017].

Cheng Wang Proposed a model are assessed on subtitle era and picture sentence recovery utilizing bidirectional LSTM. They Prove that bi-directional LSTM Perform superior to other neural system. In any case, there is a constraint of Time Consuming in Image Retrieval process additionally Language portrayal is troublesome [Cheng Wang, ACM, 2016].

Table 1: Comparative analysis of Image Captioning techniques and Model

Image captioning Technique and Model	Advantage	Limitations
Hidden Markov Model question [Alex Graves, arxiv, 2005]	Using duration weighted error, increased its recognition accuracy	Using duration weighted error, decreased the classification Performance of BLSTM
Structured max-margin objective [Jeff Donahue, Pattern analysis,2016]	Used for object detection method helps predictions for the image sentence retrieval task by using fragmentation	Phrases describe some particular number of visual entities are not Modeled
Image regions based model on weak labels [Andrej Karpathy, arxiv,2014]	Generate a description of one input array of pixels at a fixed resolution	Perform on high resolution pixel required Large dataset and taking image from dataset is time consuming task
Visual attention model & backpropagation techniques [kelvin xu,arxiv,2016].	Use for attending correct object and Using this technique perform both two-way direction operation	difficulty to attending more than one object And Identification of object require visual attention
Bidirectional Beam Search [Qing sun,IEEE,2017]	Fill-in-the-Blank Image Captioning task which used both past and future sentence structure to reconstruct sensible image descriptions.	Time Consuming To select proper Description in fill the blanks and to identify proper Path to fill the blank is difficult
bidirectional LSTM [Cheng Wang, ACM, 2016] Alex Graves, arxiv, 2005]	Bi-directional LSTM Perform better than another neural network	Time Consuming in Image Retrieval process and Language representation is difficult
Long Short-Term Memory [Alex Graves, arxiv, 2005]	Much faster and also more accurate than both standard Recurrent Neural Nets (RNNs) and time-windowed Multilayer Perceptron's	Hidden variable which are use for selection process is not identified.so that variable is used for identify an object
image guides visual and captions generator [Jonghwan Mun, arxiv,2016]	guide image in proper way and generate proper image sentences using guidance capture generator	Sometimes Captures objects not described in guidance caption also many objects are identified so it's difficult to manage them

3. Neural Network

3.1 Artificial Neural Network

A manufactured neural system is made out of numerous simulated neurons that are connected together as per a particular system design. The goal of the neural system is to change the contributions to significant outputs.

There are three layers in ANN

1) Input layer
2) Hidden layer
3) Output layer

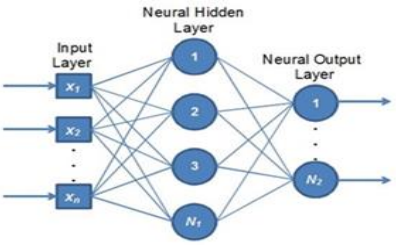


Fig.1 Artificial Neural Network

Undertakings to be understood by manufactured neural systems:

1) Controlling the developments visual data
2) Recognizing a visual question (e.g., a natural face)

3.2 LSTM (Long Short-Term Memory)

Long short-term memory (LSTM) is Stored esteems are not altered as learning continues. RNNs permit forward and in reverse associations between neurons. To order the procedure and anticipate time arrangement given time is perform by LSTM. Relative heartlessness to hole length gives favourable position to LSTM over option RNNs, shrouded Markov models and other arrangement learning strategies in various applications.

Recurrent neural networks are proficient to catching long-remove conditions, however they bomb because of the angle vanishing/detonating issues. LSTMs is acquaint for with take care of RNNs inclination vanishing issues.

LSTM systems present another structure called a memory cell.

- Every memory cell contains four primary components:
-
    - Input gate
    - Forget gate
    - Output gate
    - Neuron with a self-recurrent

These gates allow the cells to keep and access information over long periods of time. Figure 2 gives the basic structure of an LSTM unit [kelvin xu, arxiv, 2016].

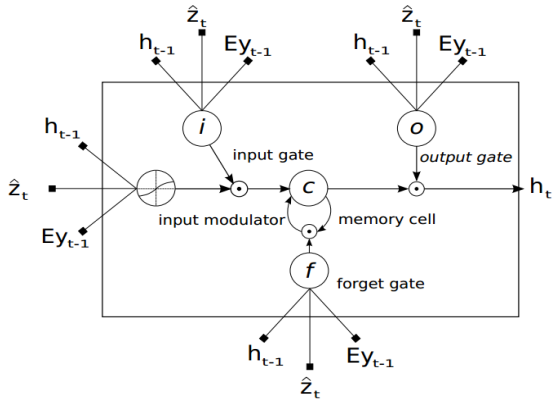


Fig.2 LSTM Architecture

where Zt is the info vector like word inserting at time t, ht is the shrouded state called yield vector putting away all the valuable data at and before time t. σ is the component astute sigmoid capacity and is the component savvy item. Ui , Uf , Uc, Uo indicate the weight networks of various entryways for input Zt , and Wi ,Wf ,Wc,Wo are the weight frameworks for concealed state ht . bi , bf , bc, bo signify the inclination vectors.

3.3 CNN

Convolutional neural network (CNN, or ConvNe) is a class of deep, feed-forward artificial neural networks that utilized for convolutional and pooling purpose. To lessen assets for picture retrieval, process a Convolutional neural system was present. There are four layer of CNN:

1) Input layer
2) Fully associated layer
3) Convolution
4) Classifier and polling layer

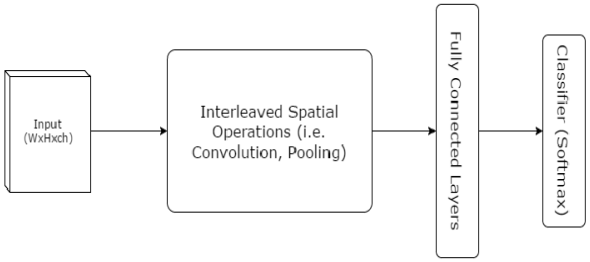


Fig.3. CNN Architecture

Initially, CNN taking network made picture as an input. Then its play out a two operation. Named by Convolution and Pooling. The target of these operations is to learn different spatial data of pictures. In the first place the convolution operation performs 2D convolution. however, the critical operation of CNN is edge recognition. At that point the pooling operation fills two fundamental needs. It lessens the extent of yield by decrease assets from accessible data. its, additionally perform interpretation and turn of system.

3.4. Bi-Directional LSTM

With a specific end goal to make utilization of both the past and future setting data of a sentence in foreseeing word, in this way, bidirectional model by sustaining sentence to LSTM from forward and in reverse request .

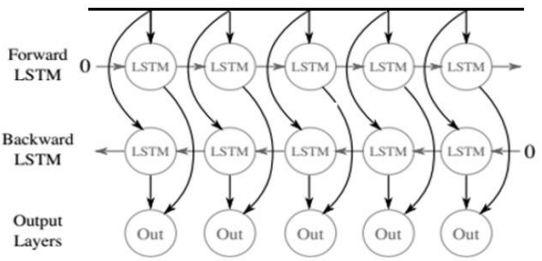


Fig.4 Bi-Directional LSTM Architecture

Figure 4: A model outlines setting data from both left and right side for producing sentence word by word after some time utilizing that sustain sentence in both forward in reverse request. For limit a join less this model is prepared by end-to-end association.

4. Proposed Model

There are three modules in this approach:

1) Input (Image)
2) Convolutional Neural Network with Bidirectional LSTM
3) Output (Generate Descriptive Sentences from Image)

This is a Hybrid Approach of Bi directional LSTM with CNN. The System Taking Input as an Image and Extract Features of Image By using Convolutional Neural Network. Feature extraction involves reducing the amount of resources required to describe a large set of data. So, basically its use for increase performance of System. Now, edge detect process will take a place that detect edges of image and finally, pooling operation used for reduce overload or output from CNN.

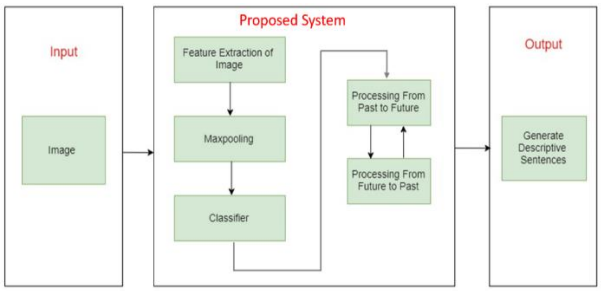


Fig.5 Bi-Directional LSTM with CNN Model

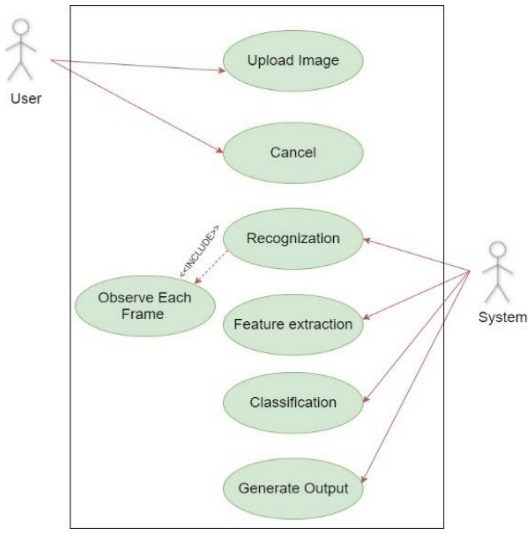
Now, Bi-directional Taking Input from the CNN and performing both Forward and Backward Direction. Forward LSTM work on past to future direction and backward LSTM work on future to past direction from available dataset. Combination of CNN and Bidirectional produce descriptive sentences of image.

5. Diagrams

5.1 Usecase Diagram

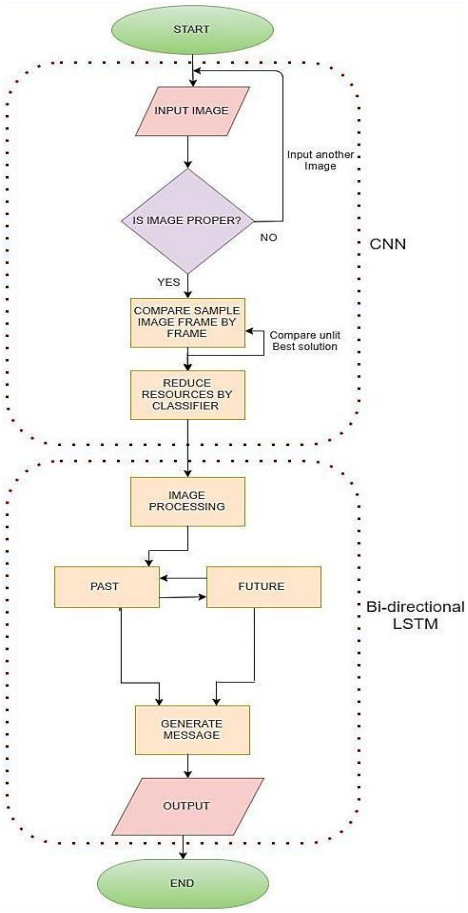
In this Usecase diagram there is one user and one system is there. User can upload image and cancel image. System can be done in four parts:

1) Recognition of an image
2) Feature Extraction of image
3) Classification of image
4) Generate output



5.2 System Flow Chart

In this Flowchart There are Start and End State. There are two part of this flowchart one is CNN and another is Bi directional LSTM. There are two process are there one is input and second one is descriptive sentences of image.one decision is also take a position in flow chart.

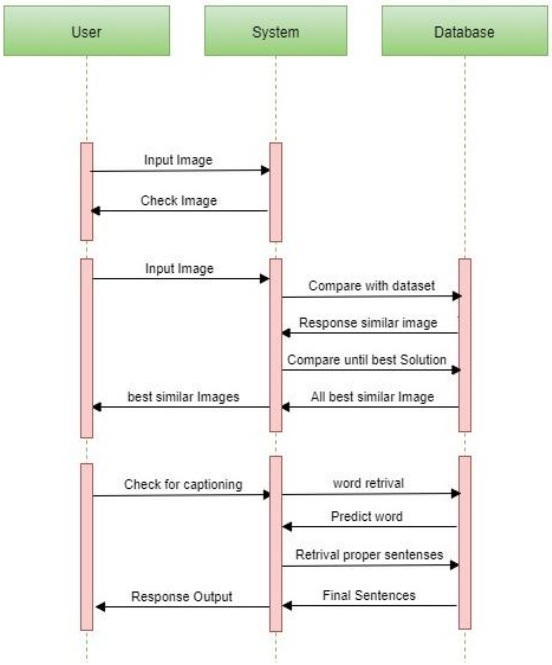


5.3 Sequence Diagram

This is a sub-part of this sequence diagram.

1) User

2) System
3) Database



Conclusions

Bi-directional LSTM model generates descriptive sentence for image, considering both past and future context. CNN for feature extraction reduces the amount of resources required to describe large set of data. Proposed approach combines CNN for feature extraction with Bi directional LSTM model to reduce the amount of resources required to generate descriptive sentence for image. In future we implement identify word or sentences from image using bi-directional LSTM on Image captioning.

References

Cheng Wang, Haojin Yang, Christian Bartz, Christoph Meinel[2016], Image Captioning with Deep Bidirectional LSTMs , ACM on Multimedia Conference, Amsterdam, The Netherlands — October 15 – 19.

Alex Graves and Jurgen Schmidhuber[2005], Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures, Neural Network. Arxiv, Jun-Jul;18(5-6):602-10.

Qing sun,Stefan Virginia, Dhruv Georgia [2017], Bidirectional Beam Search: Forward-Backward Inference in Neural Sequence Models for Fill-in-the-Blank Image, Captioning Published at IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017

Jonghwan Mun, Minsu Cho, Bohyung Han[2016] ,Text-guided Attention Model for Image Captioning, Department of Computer Science and Engineering, ARXIV Publication, POSTECH, Korea

Oriol Vinyals, Alexander Toshev ,Samy Bengio, DumitruErhan [2015], Show, Attend and Tell: Neural Image Caption Generation, IEEE ,Provided by Computer Vision Foundation.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho[2016], Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, arXiv:1502 publication.

Andrej Karpathy, Armand Joulin and Li Fei-Fei[2015], Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, Department of Computer Science, ACM Publication, Stanford University, Stanford, CA 94305, USA,

Alex Graves, Santiago Fernandez, Jurgen Schmidhuber [2005], Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition, International Conference on Artificial Neural Networks, ICANN

Andrej Karpathy, Li Fei-Fei[2015], Deep Visual-Semantic Alignments for Generating Image Descriptions, Transactions on Pattern Analysis and Machine Intelligence journal of latex class files, vol. 14, NO. 8, august.

Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama,Kate Saenko, Trevor Darrell [2016], Long-term Recurrent Convolutional Networks for Visual Recognition and Description , Arxiv, Transactions on Pattern Analysis and Machine Intelligence

Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu and Heng [2017], Attention-based LSTM and Semantic Consistency, IEEE transaction on Multimedia

https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

Jason P.C. Chiu, Eric Nichols [2016], Named Entity Recognition with Bidirectional LSTM-CNNs, Transactions of the Association for Computational Linguistics, vol. 4, pp. 357–370,

Jason P.C. Chiu, Eric Nichols [2016], Sequential Labeling with Bidirectional LSTM-CNNs, The Association for natural language processing.

Bryan A. Plummer , Liwei Wang , Chris M. Cervantes , Juan C. Caicedo[2015] ,Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models, IEEE CVPR2105,Provided by Computer Vision Foundation,.

Lin Ma Zhengdong Lu Hang Li[2015], Learning to Answer Questions From Image Using Convolutional Neural Network, Arxiv Association for the Advancement of Artificial Intelligence

Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, Ruslan Salakhutdinov[2015], Learning to Answer Questions From Image Using Convolutional Neural Network", Arxiv publication,Sep 25.

Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell[2016], Long-term Recurrent Convolutional Networks for visual Recognition and Description,Arxiv, Transactions on Pattern Analysis and Machine Intelligence

Andrej Karpathy, Armand Joulin, Li Fei-Fei[2014], Deep Fragment Embeddings for Bidirectional Image Sentence Mapping" Arxiv,22 june, Computer Vision and Pattern Recognition

https://en.wikipedia.org/wiki/Convolutional_neural_network

http://cs231n.github.io/convolutional-networks/

http://www.aclweb.org/anthology/P16-11

http://www.thushv.com/computer_vision/convolutional-neural-networks-mayor-of-the-visionville/