A Systematic Literature Review of Multi Modal Learning With Deep Neural Networks

D.Jagadish
Department of computer science & Engineering
Anil Neerukonda Institute of Technology and Sciences
Visakhapatnam, Andhra Pradesh, India

Dr.K.Selvani Deepthi
Associate Professor
Department of computer science & Engineering
Anil Neerukonda Institute of Technology and Sciences
Visakhapatnam , Andhra Pradesh , India

P.Himakar
Department of computer science & Engineering
Anil Neerukonda Institute of Technology and Sciences
Visakhapatnam, Andhra Pradesh, India

J.Ramya
Department of computer science & Engineering
Anil Neerukonda Institute of Technology and Sciences
Visakhapatnam, Andhra Pradesh, India

E.Sai teja
Department of computer science & Engineering
Anil Neerukonda Institute of Technology and Sciences
Visakhapatnam, Andhra Pradesh, India

Abstract- This paper shows the survey of the methods which are used for generation an image captioning. The approaches attach datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data. Basically it is an model based on a thorough combining of Cnn over image regions, bidirectional Rnn over sentences, and an objective which was structured that regulates the two modals through a multimodal embedding .Then it constures a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. There are various parameters are associated like data sets, models used to generate the descriptions.

Keywords – Neural Network (CNN), Recurrent Neural Network (RNN), data set.

I. INTRODUCTION

A keen staring towards an picture is enough for a person's brain to explain an great deal of details about the visual scene. to automatically describe the content of a picture using properly formed English sentences may be a very tough task ,although it will have impact , for example by helping visually impaired people to know image. This task is taken into account as hard to realize , for instance ,the well-studied image classification or visual perception tasks, which are a main focus within the computer vision. Description must

capture not only the objects contained in a picture , but it should define how these objects relates to each other and to their attributes and therefore the activities they're involved within the image . Moreover, the above semantics has got to be expressed during a tongue like English, which suggests that a language model is required additionally to visual understanding. the most inspiration of this work comes from recent improvements in deep learning, where the task is to rework a sentence written during a language , into its translation T within the target language. From last decade, it had been achieved by a separate & different tasks so as , but recent work has shown that translation are often wiped out a way simpler way using Neural Networks such as recurrent neural networks (rnn's),convolutional neural networks.

1.1 Convolutional Neural Network:

Convolutional neural network is a class of neural network which is widely used in visual imagery. Image captioning task is split into two modules – one is a picture based model and another language based model. In this model, the features of image are extracted, they are translated in language model. For image based model, CNN is employed and RNN is employed for laguage based model. the subsequent figure summarises this approach:

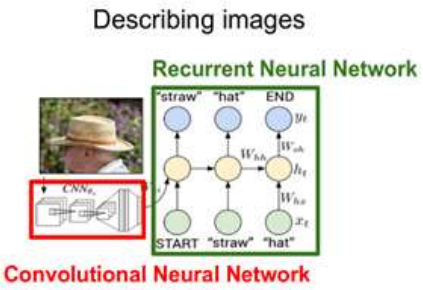

Figure 1:cnn as encoder & rnn as decoder

CNN which acts as an encoder extracts the features of the image. to coach LSTM which acts as decoder one should predefine label and target text. Label is nothing but identifying all the words within the caption having at the start and target is same as label with no .Instead of at the start , there's at the ending of labelled sequence.Typical Convolution Neural specification consists of several convolutional layers (each of various size) and pooling layers alternatively, followed by a couple of fully connected layers and soft-max layer. In fully connected layers, each neuron in current layer is connected to all or any the neurons within the previous layer but the neurons in convolutional layer receives the input from those in their receptive field which ends up in extraction of Features of a picture is extracted from fully connected layer of the VGG16 network. Neurons which are present in convolutional layer are organized in feature maps and every one the neurons carry same weights and perform same operation in several parts of a picture . CNN module consists of 4 convolutions (64, 128, 256and 512). The output of the CNN module may be a 512- dimensions vector for every word . the next layer is pooling . the idea behind a pooling layer is to gather together features from maps generated by convolving a filter over a picture . then the function scale back the spatial size of the representation and to the amount of parameters within the network. There are two sorts of pooling – one is average pooling and another one is max pooling. Max pooling are often done by applying a max filter to the non-overlapping regions.

1.2 Recurrent Neural Network:

The extracted features from CNN becomes input to the RNN. during this sort of neural network, the output from the previous layer becomes input to the present layer. The important feature of RNN is its hidden state which stores some information about the layers. allow us to consider (x1…….xt) as inputs and RNN generates (y1……..yt) as output by applying the below recurrence formula

$$h_t=\emptyset(h_{(t-1)},x_t) \rightarrow \text{equation-1}$$

Where ht is internal hidden state at time step t, ϕ is non-linear activation function Ht-1 is hidden state of previous time step

The non-linear activation function could also be a sigmoid function or hyperbolic tangent function. The sigmoid function is denoted by σ and tangential function is represented as tanh. Applying elementwise tanh to the above equation:

$$h_t=\tanh(w_{hh} h_{(t-1)}+ w_{xh} x_t+b_h) \rightarrow \text{equation-2}$$

Where Whh and Wxh are learned weight matrices.
Output are often obtained by $y_t=w_{hy} h_{(t)}+b_y$, by is that the bias.

1.3 VGG16:

Visual Geometry Group (VGG) may be a convolutional neural network model that's trained on quite million images from ImageNet dataset.It adds ReLU and use a linear layer to get 512-dimensional embedding.

1.4 Data Sets:

Data sets is a collection of instances which we use them to feed & train our models.

1.5 Ms coco:

This dataset doesn't specialise in iconic views that contain only one view of one object such objects within the background, partially occluded and amidst clutter also are present and are important for image retrieval tasks. Iconic images contain one subject within the image frame that's clearly defined and occupies most space within the image frame.That type of images are completely or easily recognized. Detailed spatial understanding of the thing layout may be a core component of scene analysis. An objects spatial location are often defined coarsely employing a bounding box or with precise pixel level segmentations.

1.5.1Images features:

- Pair of objects in conjunction with images retrieved using scene based queries.[15]
- Labeling is completed because the image containing a specific objet category using hierarchal labelling.[15]
- Individual instances are labeled and verified and eventually segmented.[15]

1.5.2 Images count:

- 91 common object categories with 82 having quite 50000 labelled instances.[15]
- 2,500,000 instances in 328,000 images.[15]
- Number of labeled instances per image help in learning contextual information.[15]

1.6 Flickr 30k:

The Flickr30k dataset has become a typical benchmark for sentence-based image description. Flickr30k Entities, which augments the 158k captions from Flickr30k with 244k coreference chains, linking mentions of an equivalent entities across different captions for an equivalent image, and associating them with 276k manually annotated bounding boxes. Such annotations are important for continue progress in automatic image description and understanding. they allow us to define a replacement benchmark for localization of textual entity mentions in a picture . We present a robust baseline for this task that mixes an image-text embedding, detectors for common objects, a color classifier, and a bias towards selecting larger objects. While our baseline rivals in accuracy more complex state-of-the-art models, we show that its gains can't be easily parlayed into improvements on such tasks as image-sentence retrieval, thus underlining the restrictions of current methods and therefore the need for further research.

II. APPROACHES

2.1 Multimodal Space:

Multimodal space-based method consists of language Encoder, vision, multimodal space, and language decoder. The vision uses a cnn to extract the features of images. language encoder extracts the word features and learns a dense feature corresponding of each word. It then forwards the semantic context to there current layers. Then multimodal space maps the image features into a common space with the word features.

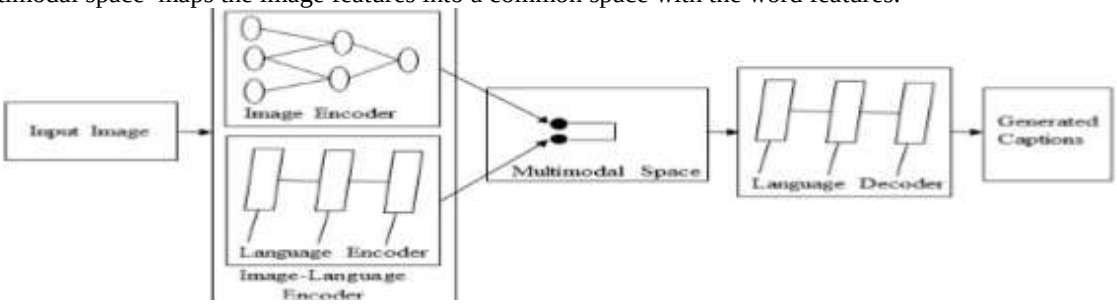

Figure 2:A block diagram of multimodal space-based image captioning.

The resulting map is then feeded to language decoder that generates captions by decoding the map. Deep neural networks and multimodal neural language model are used to learn both image and text combinedly in a multimodal space.

2.2 Dense Captioning:

Dense Captioning method localizes all the salient regions of an image and then it generates descriptions for those regions[12].These Region proposals are generated for the different type of regions in the image.CNN is used to obtain the region-based image features.the outputs of cnn are used by a language model to generate captions for every region.

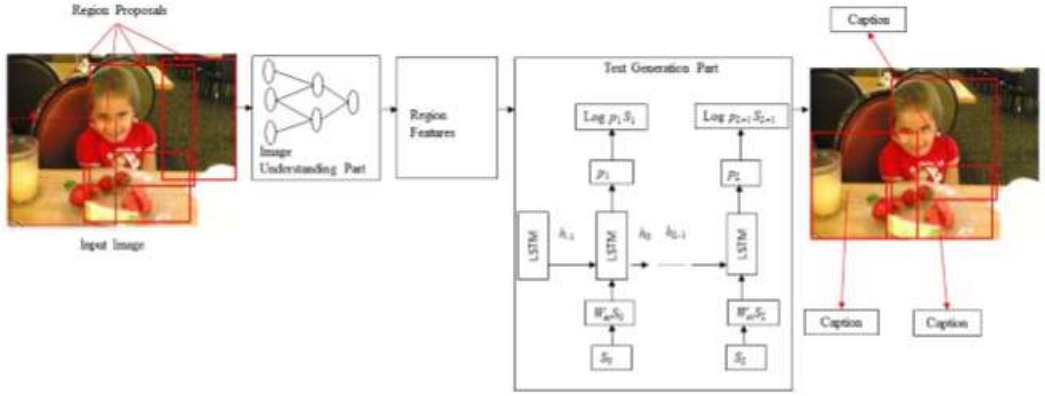

Figure 3:A block diagram of a typical dense captioning method.

Dense captioning proposes a fully convolutional localization network architecture, which is consists of a convolutional network ,dense localization layer, and LSTM language model . The dense localization layer processes an image with a single ,efficient forward pass, which predicts a set of region of interest in the image. Therefore ,it requires no external region proposals unlike to Fast the R-CNN. The working principle of the localization layer is related to the work of Faster R-CNN .However, Johnson et al uses a differential ,spatial soft attention mechanism. This modification helps the method to back propagate through the network and smoothly select the active regions .Visual Genome dataset is used for the experiments in generating region level image captions.

2.3 Encoder-Decoder Architecture-Based Image captioning:

The neural network-based image captioning methods similar to the encoder-decoder frame work-based neural machine translation .In this approaches ,global image features are extracted from the hidden activations of CNN and

then fed them in to an LSTM to generate a sequence of words. A typical method of this category has the following general steps:
(1) Avanilla CNN is used to obtain the scene type,to detect the objects and their relationships.
(2) The output of above process is used by a language model to convert them into words, combined phrases that produce an image captions.

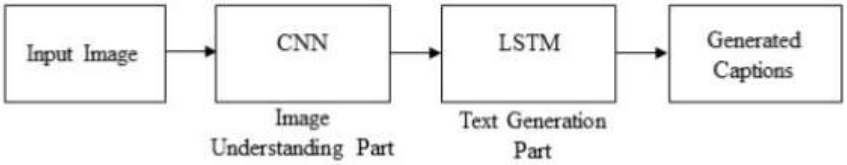

Figure 4:A block diagram of simple Encoder-Decoder architecture-based image captioning

Neural Image Caption Generator(NIC) method uses a CNN for image representations and an LSTM for generating image captions.CNN uses a innovative method for batch normalization and the output of the last hidden layer of CNN is used as an input to the LSTM decoder. LSTM is capable of keeping archive of the objects that already have been described using text.NIC is tained based on maximum likelihood estimation.

2.4 Attention based Image Captioning:

Attention based mechanisms are becoming increasingly popular in deep learning because they can address these limitations. They can focus on the various parts of the input image while the output sequences are being produced. Image information is obtained based on the whole scene by a CNN.The language generation part generates words or phrases based on the output of cnn.Salient regions of the given image are focused in each time step of the language generation model based on generated words or phrases.Captions are updated dynamically until the end state of language generation model.

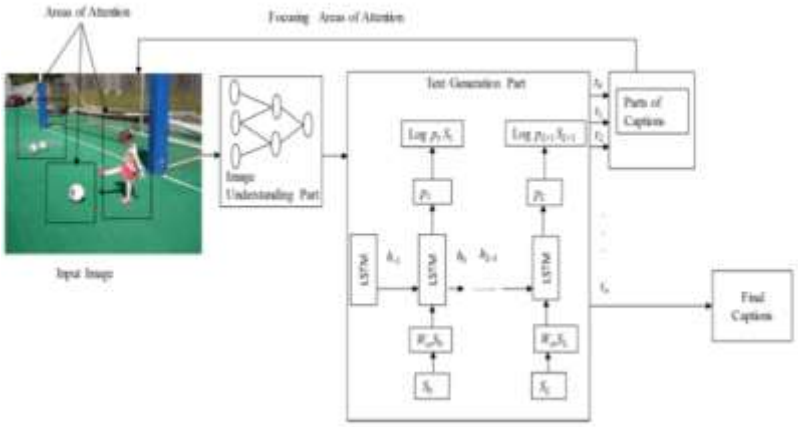

Figure 5:A block diagram of the attention-based image captioning method

III.LITERATURE SURVEY:

3.1 Existing Methods:

3.1.1.Sharmistha Jat et al[2019] proposed Relating Simple Sentence Representations in Deep Neural Networks and the Brain. The main Objective is studied relationship between sentence representations learned by deep neural network models and those encoded by the brain. Algorithm used is Random Embedding Model, GloVe Additive Embedding Model, Simple Bi-directional LSTM Language Model. The major Limitation is the shallower layers represent low-level features and the deeper layers represent more task-oriented features.

3.1.2**.**Jyoti Aneja et al proposed Convolutional Image Captioning.The main Objective is Image captioning, i.e., describing the content observed in an image, with attention mechanism. Algorithm used is Rnn approach ,feed-forward Cnn approach. The major Limitation is RNNs tend to produce lower classification accuracy, and, despite LSTM units, they still suffer to some degree from vanishing gradients.

3.1.3**.**Tsung-Yi Lin et al[2015] proposed Microsoft COCO: Common Objects in Context. The main Objective is to provide an dataset with the goal of advancing the state-of-the-art in object recognition. Algorithm used is Bounding box ,Segmentation detection algorithm. The major Limitation is the predicted segmentation might not recover object detail even though detection and ground truth bounding boxes overlap well.

3.1. 4.Zhongliang Yang et al[2017] proposed Image Captioning with Object Detection and Localization. The main Objective is to Automatically generating a natural language description of an image. Algorithm used is Deep recurrent neural networks with lstm & attention generation. The major Limitation is features can't extract from the image without the spatial information.

3.1.5**.**Andrej Karpathy et al[2015] proposed Deep Visual-Semantic Alignments for Generating Image Descriptions. The main Objective to present a model that generates natural language descriptions of images and their regions. Algorithm used is Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences. The major Limitation is the RNN receives the image information only through additive bias interactions, which are known to be less expressive than more complicated multiplicative interactions.

3.1.6.OlgaRussakovsky et al[2015] proposed ImageNet Large Scale Visual Recognition challenge. The main Objectives are to discuss the challenges of creating this large-scale object recognition benchmark data. To highlight the developments in object classification and detection that have resulted from this effort. To take a closer look at the current state of the field of categorical object recognition. Algorithm used is ground truth bounding box algorithm. The major Limitation is the images contains the filters which are enhanced by the humans produces incorrect annotations.

3.1.7.Shuang Bai et al[2018]. proposed A Survey on Automatic Image Caption Generation.The main Objectives is to classify image captioning approaches into different categories. Representative methods in each category are summarized. Algorithm used is Retrieval based image captioning, Template based image captioning. The major Limitation in Retrieval based methods is having large limitations to their capability to describe images. using rigid templates as main structures of sentences will make generated descriptions less natural.

3.1.8.Moses Soh et al[2016] proposed Learning CNN-LSTM Architectures for Image Caption Generation. The main Objectives is to Implements a generative CNN-LSTM model that beats human baselines by 2.7 BLEU-4 points**.** Algorithm used is Convolutional neural networks(cnns) with lstm model, BLEU. The major Limitation is Partial errors tend to occur due to lack of attention to specific details in images.

3.1.9.Lakshmi narasimhan Srinivasan et al[2018] proposed Image Captioning - A Deep Learning Approach**.** The main Objectives is Propose a hybrid system employing the use of multilayer Convolutional Neural Network (CNN) to generate vocabulary describing the images and a Long Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. Algorithm used is Encoder-decoder model, Convolutional Neural Networks(cnn's),Recurrent neural networks(rnn's) with

lstm. The limitations of neural networks are determined mostly by the amount of memory available on the GPUs used to train the network as well as the duration of training time it is allowed.

3.1.10**.**Pranay et al[2017] proposed Camera2Caption:A Real-Time Image Caption Generator. The main Objectives is to Present our simplistic encoder and decoder based implementation with significant modifications and optimizations which enable us to run these models on low-end hardware of hand-held devices. Simplest Encoder-decoder model & long short term memory[Lstm]. Algorithm used is The major Limitation . This model doesn't use visual attention, so efficiency is low for the generated captions.

IV.EVALUATION METRICS:

4.1 Bleu:

Bilingual Evaluation Understudy could be a measurement of the differences between an automatic translation and one or more human-created reference translations of an equivalent source sentence. BLEU's evaluation system requires two inputs:

(i) a numerical translation closeness metric, which is then assigned and measured against human created references.
(ii) a corpus of human reference translations. BLEU averages out various metrics using an n Gram's method ,probabilistic language model often utilized in linguistics.

The result's typically measured on a 0 to 1 scale, with 1 because the hypothetical "perfect" translation. Since the human reference, against which MT is measured, is usually made from multiple translations, even a person's translation wouldn't score a 1, however. Sometimes the score is expressed as multiplied by 100 or,as in the case of mentioned above,by10.

4.1.2 Rouge:

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially of a group of metrics for evaluating automatic summarization of texts also as MT . It works by comparing an automatically produced summary or translation against a set of reference summaries .The results of given two equations are compared to evaluate the sentences.

X1=O/rs →equation-3

X1=comparison of reference summary with o.
o=number of overlapping words with reference summary
rs=total words in reference summary

X2=O2/SS →equation-4

X2=comparison of system summary with o
O2=number of overlapping words with system summary
SS=total words in system summary

ROUGE-N, ROUGE-S and ROUGE-L are often thought of because the granularity of texts being compared between the system summaries and reference summaries.

V.CONCLUSION:

Image captioning may be a very exciting exercise and raises tough competition among researchers. There are more and more scientists who are deciding to explore this study field , therefore the amount of data is consistently

increasing. it had been noticed that the results are usually compared with quite old articles, although there are dozens of latest ones, it requires an more research to urge the higher results and new ideas for improvements. this systematic literature review summarizes all the most recent articles and their results in one place. The amount of knowledge will never stop increasing and new information will keep appearing, so future studies should consider if static models are good enough for there requirements.

REFERENCES:

[1] Ingrid Hrga, Juraj Dobrila "Deep Image Captioning:Models, Data and Evaluation", Dobrila University of Pula

[2] Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang, "Image Captioning with Object Detection and Localization", Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

[3] Andrej Karpathy, Li Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions", Stanford university.

[4] O. Russakovsky, J. Deng, H,Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.Berstein, A. C. Berg, and L.Fei-Fei. "Imagenet large scale visual recognition challenge. International Journal of Computer Vision", 115(3):211-252, Dec 2015

[5] A Survey on Automatic Image Caption Generation Shuang ,Bai_School of Electronic and Information Engineering, Beijing Jiaotong University, No.3 Shang Yuan Cun, Hai Dian District, Beijing, China

[6] Moses Soh "Learning CNN-LSTM architectures for Image Caption Generation" Department of Computer Science, Stanford University

[7] Lakshminarasimhan Srinivasan1, Dinesh Sreekanthan2, Amrutha A.L3 "Image Captioning - A Deep Learning Approach".

[8] JyotiAneja∗, Aditya Deshpande∗, Alexander G. Schwing, "Convolutional Image Captioning",University of Illinois at Urbana-Champaign

[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, RuslanSalakhutdinov, Richard Zemel, YoshuaBengio "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". ICML 2015.

[10] .ZAKIR HOSSAIN, FERDOUS SOHEL "A Comprehensive Survey of Deep Learning for Image Captioning" MD,MurdochUniversity,Australia.

[11] Raimonda Stani¯ut˙e, "A Systematic Literature Review on Image Captioning".

[12] https://www.cs.princeton.edu/courses/archive/spring18/cos598B/public/outline/Towards%20image%20captioning.pdf

[13] http://sidgan.me/technical/2016/01/09/Exploring-Datasets

[14] https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2

[15] https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/