International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 01 | Jan 2021 www.irjet.net p-ISSN: 2395-0072

Real Time Image Captioning and Voice Synthesis using Neural Network

Jasmita Khatal[1], Prajkta Jadhav[2], Shraddha Parab[3], Prof. Rasika Shintre[4]

[1,2,3]*Student, Computer Engineering, SIGCE, Navi Mumbai, Maharashtra, India*
[4]*Asst. Professor, Computer Engineering, Smt. Indira Gandhi College of Engineering, Navi Mumbai, Maharashtra, India*

Abstract –*In Today's world there are lot of images are captured using numerous of devices in Mobile Camera which contains different objects. But if someone captured a image in their Smartphone but not able to recognize what image is showing, so in this case Image Captioning will be useful to get final output of the image, If any user have disabilities like visually impaired then these technology might help them to understand and recognize the Image properly. This project is all about Real Time Image Captioning with Voice Synthesis would help user to get final output. First understand about Image Caption generator. This Function involves computer Vision and Natural Language Processing concept to recognize the context of image and describing them in a Natural Language like English. This Constant Increase is due to the ease of capturing image Service in Portable device such as Mobile, Tablets, I-pad or small Camera. We are using general Neural Network configuration that combine two Supervisory Signals that is Image based text-captions and Text based speech at output, in the training phase and generate captions for given images in first phase. This project has main idea that the image caption (text) and help the Real time image captioning model learn to focus on important frames.*

***Key Words*: Natural language processing (NLP), computer vision, CNN, RNN, LSTM, Real-time, voice synthesis**

1. INTRODUCTION

Images are important part of human life. People love to create memories by taking their Pictures from Different cameras. Images are not only used for to create Memories but also giving us important information. Images Helped small children to understand different structures to understand colours. Now there is Rapid Growth in the technology, and Spread of internet all over the world there is lot of datasets which are Present in the Internet.

Deep learning often won't to automatically annotate these images, thus replacing the manual annotations done. This may greatly reduce the human error also because the efforts by removing the necessity for human intervention. The sector brings together state-of-the-art models in tongue Processing and Computer Vision, two of the main fields in AI. One among the challenges is availability of huge number of images with their associated text ever expanding internet. However, most of this data is noisy and hence it can't be directly utilized in image captioning model. For training a picture caption generation model, an enormous dataset with properly available annotated image is required. In these worlds images are generated by Human Intervention, it become an impossible task to get bug Commercial Datasets. Those images are stored into the Databases. The Image Databases has given an input to Deep Learning neural Network CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network).

Captioning is very Important It Need the Computer Vision Technology which is very useful as well as it Needs Natural Language Processing. Both Technologies are used to get accurate Sentences. Objects are get extracted using by CNN which are convolutional Neural Networks these are passed to Language Model.

1.1 OBJECTIVE

1. Understanding the concept of Convolution Neural Networks (CNN) for feature extraction and feature vector. Understanding the visual semantics in the real time image and converting it into a simple (partial) caption.
2. Image caption, automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing.
3. The application of image caption is extensive and significant, for example, the realization of human-computer interaction.
4. Less storage space required.
5. Understanding the concept of Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM).

1.2 SCOPE

The World is moving towards digital. Lots of images are uploaded everyday on Internal/cloud. There are numerous dataset are available on internet. Real time captioning as well as Voice synthesis is first time implemented. For using of this, there are some condition where understanding image and what that image giving

the information is sometimes not easy to get. So this helps to persons who have Disability (Visually Impaired).

1.3 FEATURES

1. Users get output in Real-time.
2. The Image captions are accurate compared to others Methods.

2. LITERATURE SURVEY

The image captioning system has become more popular in day by day. Neural network makes it efficient and accessible to everywhere. As per Kaustubh Shivdikar and Kshitij Marwah in [1] introduces hybrid engine methods. Hybrid engine utilizes the combination of "Speed Up Robust Algorithm" (SURF) with minimum eigen value to notice and classify objects. Then passed to a Content Free Grammar (CFG) to create grammatically meaningful phrases. The detection of the object is performed on the dataset and uncaptioned images. Primary feature description is performed using the SURF algorithm. Here Hessen matrix is used in SURF based object detection.

In [2], research paper of Kun Fu, Junqi Jin, Runpeng Cui, Fei Sha, Changshui Zhang developed captioning framework that exploits parallel structures between sentences and images. In this method, there is a nearby correspondence between visual ideas that recognize object regions and their corresponding sentences. In addition, the procedure of creating the next word, given the already created ones, is lined up with the visual observation experience where the consideration moving among the areas forces a requesting of visual recognition preference.

In [3], K.C. Nithya and Vinod Kumar focus on region and scene specific context, parallel-fusion, cascade attention, two-phase learning methods for image captioning. The problems are to get proper information from image captioning means textual depiction of image. To get proper information from image captioning means textual depiction of image. Captioning with correct sentences needs computer vision and Natural language processing (NLP) for getting correct sentences. Classified objects are then passed to language model to make captions. Semantic knowledge about an object in a picture need to obtain by capturing characteristics of an image globally and locally.

In [4], Daniel Thirman a model which is used to generate novel image captions for a previously unseen image by using a combination of a recurrent neural network and a convolutional neural network. Some of these include image recognition for autonomous robotic systems which need to be able to recognize and process what they see, creating an image database that is search-able by keywords without having to manually tag and describe the different images that are added to the database, as well as many other possible applications.

In [5], C. Khancome, V. Boonjing, and P. Chanvarasuth-Generate novel image captions for a previously unseen image by using a combination of a recurrent neural network and a convolutional neural network. Model trained on Flicker datasets and networks are used to get How perplexity how good language model and Proper sentence given in an image. The result was accurate and proper sentences are generated.

3. PROBLEM STATEMENT

In Today's life lot of images are uploaded on internet. Images contain valuable information. Some time that information not easily get understood for that type of problem Image captioning is very useful. Our research problem is to accurately identify objects using neural networks and after using Computer vision and Natural Language Processing, language models captions are generated.

4. PROPOSED SYSTEM

In our proposed system, Deep learning are often wont to automatically annotate these images, thus replacing the manual annotations done. this may greatly reduce the human error also because the efforts by removing the necessity for human intervention The generation of captions from images has various practical benefits, starting from aiding the visually impaired, to enabling the automated, cost-saving labelling of the many images uploaded to the web a day, recommendations in editing applications, beneficial in virtual assistants, for indexing of images, for dim-sighted people, for social media, and a number of other other tongue processing applications. The sector brings together state-of-the-art models in tongue Processing and Computer Vision, two of the main fields in AI. One among the challenges is availability of huge number of images with their associated text ever expanding internet.

Image is captured by the System. That image is Processed by Concurrent Neural Network for the features or Objects in the Images Extracted then the Recurrent Neural Network using all the Extracted Features transformed into to Understandable Description of image.

After Using Neural Networks with using of three technologies which are computer Vision, Natural language Processing and Language Models helps to get proper and accurate Captions.

4.1 FLOWCHART

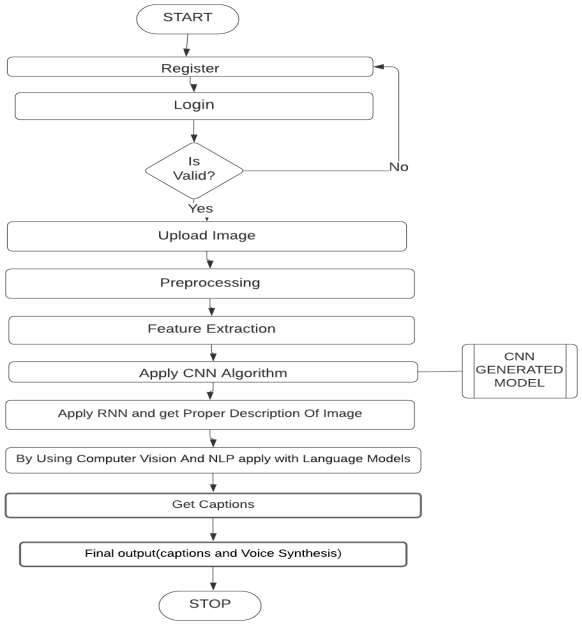

Figure 4.1.1.Flowchart

4.2 ALGORITHMS

4.2.1CNN MODEL

4.2.1.1Kernel convolution

Subsequent feature map values square measure calculated in step with the subsequent formula, wherever the input image is denoted by f and our kernel by h. The indexes of rows and columns of the result matrix square measure marked with m and n severally.

$$G[m,n] = (f * h)[m,n] = \sum_j \sum_k h[j,k] f[m-j, n-k]$$

4.2.1.2 Padding

For a gray scale (n × n) image and (f x f) filter/kernel, the size of the image ensuing from a convolution operation is (n – f + 1) x (n– f + 1).

- **Valid Padding:** It implies no padding the least bit. The input image is left in its valid/unaltered form. So, [(n x n) image] * [(f x f) filter] —> [(n – f + 1) x (n – f + 1) image]
- **Same Padding:** During this case, we tend to add 'p' padding layers specified the output image has an equivalent dimensions because the input image.
- So, [(n + 2p) x (n + 2p) image] * [(f x f) filter] —> [(n x n) image.
- Which gives p = (f – 1) / 2 (because n + 2p – f + 1 = n).

4.2.1.3 Pooling

- For a feature map having dimensions n_h x n_w x n_c the dimensions of output obtained after a pooling layer is (n_h - f + 1) / s x (n_w - f + 1)/s x n_c
- where, -n_h - height of feature map

  n_w - width of feature map

  n_c - number of channels in the feature map

  -f - size of filte

  -s- stride length

4.2.1.4 Non linearity(ReLu)

- ReLU stands for Rectified Linear Unit for a non-linear operation.
- The output is *f(x) = max(0,x).*

4.2.1.4 Fully Connected Layer

- The layer we call as FC layer, we flattened our matrix into vector and feed it into a fully connected layer like a neural network.

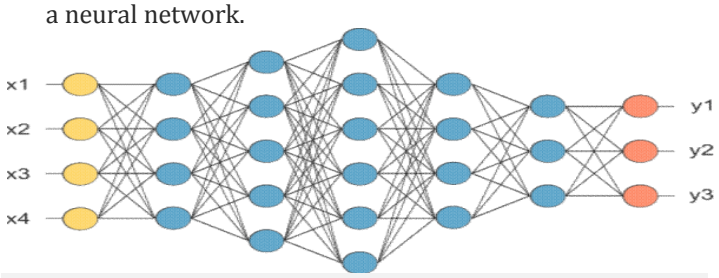

Figure 4.2.1.5.1 : After pooling layer, flattened as FC layer.

In the on top of diagram, the feature map matrix are going to be reborn as vector (x1, x2, x3 ...). With the fully connected layers, we tend to combine these options along to make a model. Finally, we've activation function like softmax to classify the outputs as cat, dog, car, truck, etc.,

4.2.2 RNN Model

- The main and most vital feature of RNN is Hidden state that remembers some info a few sequences.
- **Formula for calculating current state:**

$$h_t = f(h_{t-1}, x_t)$$

$c_t \rightarrow cell\ state(memory)\ at\ timestamp(t).$

$\tilde{c}_t \rightarrow represents\ candidate\ for\ cell\ state\ at\ timestamp(t).$

$note^*\ others\ are\ same\ as\ above.$

4.2.3 Long Short Term Memory (LSTM)

LSTM can be used to solve problems faced by the RNN model.
In LSTM we will have 3 gates:
1. Input Gate (i): It determines the extent of information to be written onto the Internal Cell State.
2. Forget Gate (f): It determines to what extent to forget the previous data.
3. Output Gate (o)
The equations for the gates in LSTM are:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$
$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$
$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

Equation of Gates

$i_t \rightarrow represents\ input\ gate.$
$f_t \rightarrow represents\ forget\ gate.$
$o_t \rightarrow represents\ output\ gate.$
$\sigma \rightarrow represents\ sigmoid\ function.$
$w_x \rightarrow weight\ for\ the\ respective\ gate(x)\ neurons.$
$h_{t-1} \rightarrow output\ of\ the\ previous\ lstm\ block(at\ timestamp\ t-1).$
$x_t \rightarrow input\ at\ current\ timestamp.$
$b_x \rightarrow biases\ for\ the\ respective\ gates(x).$

- First equation is for input gate that tells us. That what are new data we're aiming to store among the cell state(that we'll see below).
- Second is for the forget gate that tells the information to throw far from the cell state.
- Third one is for the output gate that's utilized to produce the activation to the last word output of the LSTM block at timestamp 't'.
- The equations for the cell state, candidate cell state and additionally the ultimate output:

$$\tilde{c_t} = tanh(w_c[h_{t-1}, x_t] + b_c)$$
$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t}$$
$$h_t = o_t * tanh(c^t)$$

Note: *represents the half wise multiplication of the vectors. The equations for the cell state, candidate cell state and also the final output:

Note:*represents the element wise multiplication of the vector.

4.2.4 Text to Speech (Voice) Module:

Text to speech makes associate degree system created text can scan that text (caption) and convert it into audio via speaker. Text to speech (TTS) is simple but powerful feature. The net Speech API provides two distinct areas of utility — speech recognition, and speech synthesis (also known as text to speech, or TTS) — that open up fascinating new prospects for accessibility, and management mechanisms. Text to speech conversion is completed by victimization speech genus like Google speech API.

5. CONCLUSIONS

This is an innovative project system for people and act as voice assistant for them. This system is used to visually impaired people to know with whom are taking or it is present in the surrounding by using different images and text to speech conversions. The converted voice will help them to identify the image. The Image Captioning Techniques Improving in last few years. There are various approaches or Techniques like Retrieval based Image captioning and template based Image captioning these techniques helps to get captions. But here we have used CNN (Convolutional Neural Networks) for extracting the features and RNN to get captions. But this process is not implemented in real time previously. The whole system is basically in real time image.

ACKNOWLEDGEMENT

The success and final outcome of this research required a lot of guidance and assistance and we are extremely privileged to have got this all. All that we have done is only due to such supervision and assistance and we would not forget to thank them.

We respect and thank Prof. Rasika Shintre, for providing us insight and expertise that greatly assisted the research. We are extremely thankful to her for providing such a nice support and guidance.

REFERENCES

[1] A.Karpathy, A.Joulin and L.F.Fei-Fei," Deep fragment embeddings for bidirectional image sentence mapping," In Advances in neural information processing systems,pp. 1889-1897,2014.
[2] I. Sutskever,O. Vinyals and Q. V. Le," Sequence to sequence learning with neural networks," In Advances in neural information processing systems, pp. 3104-3112,2014.
[3] A.karpathy, L.F.Fei-Fei,"Deep visual-semantic alignments for generating image descriptions," In Proceedings of the IEEE conference on computer vision and pattern recognition,, pp. 3128-3137,2014.
[4] R. Kiros, R. Salakhutdinov and R. S. Zemel," Unifying visual-semantic embeddings with multimodal neural

language models," arXiv preprint arXiv: 1411.2539,2014.
[5] J. Mao et al. "Deep captioning with multimodal recurrent neural networks (m-rnn)," arXiv preprint arXiv: 1412.6632,2014.
[6] Sun Chengjian, Songhao Zhu, Zhe Shi, "Image Annotation Via Deep Neural Network", Published in: 2015 14th IAPR International Conference on Machine Vision Applications (MVA).
[7] Venkatesh N. Murthy, SubhransuMaji, R Manmatha, "Automatic Image Annotation using Deep learning representations", ICMR '15 Proceedings of the 5th ACM on International Conference on Multimedia Retrieva.
[8] OriolVinyals, Alexander Toshev, SamyBengio, DumitruErhan, "Show and Tell: A Neural Image Caption Generator", published 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[9] Kun Fu, Junqi Jin, Runpeng Cui, FeiSha, Changshui Zhang." Aligning Where to See and What to Tell : Image Captioning with Region-based Attention and Scene-specific Contexts". IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.
[10] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," in ICLR, 2015.
[11] OriolVinyals, Alexander Toshev, SamyBengio, and DumitruErhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE Conf on Comp Vision and Pattern Rec, 2015, pp.3156–3164.
[12] JiaheShi,YaliLi, ShengjinWang."Cascade Attention: Multiple Feature Based Learning For Image Captioning".ICIP2019.

BIOGRAPHIES

Jasmita Khatal is pursuing Bachelor degree (B.E.) in Computer Engineering from Smt. Indira Gandhi College of Engineering, Navi Mumbai.

Prajkta Jadhav is pursuing Bachelor degree (B.E.) in Computer Engineering from Smt. Indira Gandhi College of Engineering, Navi Mumbai.

Shraddha Parab is pursuing Bachelor degree (B.E.) in Computer Engineering from Smt. Indira Gandhi College of Engineering, Navi Mumbai.

Prof. Rasika Shintre, Obtained the Bachelor degree (B.E. Computer) in the year 2011 from Ramrao Adik Institute of Technology (RAIT), Nerul, and Master degree (M.E. Computer) from Bharati Vidyapeeth College of Engineering, Navi Mumbai. She is Asst. Professor in Smt. Indira Gandhi College of Engineering of Mumbai university and having about 8 yrs. of experience. Her area of interest includes Data mining & information Retrieval.