Caption Recommendation System

[1]Dr.K.Spurthi, Associate Professor , CSE(AIML), Kolluspoorthy03@gmail.com
Swarna Bharathi institute of science and technology,
Khammam,
[2]B.Yugandhara Chary ,Assistant Professor, CSE(AIML), yugandhar.bandla@gmail.com
Swarna Bharathi institute of science and technology,
Khammam
[3]Ameena nasreen, Assistant Professor, CSE(AIML), amena.nasreen.md@gmail.com
Swarna Bharathi institute of science and technology,
Khammam

Abstract

A difficult neural network task is the development of a human-readable written description to a given image. This process is known as caption generation. Both knowledge of computer vision and natural language processing are necessary for this task. On social media, we see a lot of pictures every day. Images in these sources would need visitors to fill in the blanks. There are several reasons why image captioning is crucial. Facebook and Twitter, for instance, have the capability to immediately produce image-based descriptions. What we're wearing, the location (beach, café, etc.), and our activities may all be described. The ability to detect and identify items in images is crucial for automated caption generation. In addition, it must be able to recognize different types of scenes and comprehend the attributes of things as well as how they interact with one another. Syntactic and semantic knowledge of the language are necessary for sentence generation. A big number of pictures and videos are no problem for deep learning based approaches, which automatically learn characteristics from training data. We will use convolutional neural networks (CNNs) and other deep learning approaches for picture categorization, and RNN encoders and decoders for caption synthesis. Sentiment analysis and caption creation will also make use of language models like LSTM.

Keywords: Semantic, Automatic captions, Computer Vision Topics covered include sentiment analysis, title creation, deep learning, convolutional neural networks, LSTM, and backpropagation.

I. INTRODUCTION

Because there are so many people using social media, our project is centered on making captions that are suitable for the user-provided image(s). For instance, social media sites like Instagram and Facebook may deduce your location (beach, café, etc.), clothing color, and, most crucially, your activity, just by looking at your photos. It is fascinating and inspiring to see how cutting-edge technology like AI and ML can use human-provided data to analyze and understand photographs, including coming up with relevant captions. Along with this, picture captioning has intriguing uses in Picasa, Tesla (Google's self-driving vehicles), SkinVision (which can determine whether a skin issue is cancerous), Google photographs (which can categorize your photographs into various areas, such as mountains, sea, etc.), and many more. Additionally, it might be a great assistance to those who are visually challenged. One part of the suggested approach is using deep learning for picture captioning. Deep convolutional neural networks provide superior discriminative expression capability in complicated contexts compared to traditional detectors that rely on hand-crafted descriptors. These networks automatically learn from training data and build hierarchical feature representations, starting with raw pixels and progressing to high-level semantic information. We have covered the RNN-LSTM method for caption synthesis and the Faster R-CNN method for object recognition. In order to generate captions, our suggested system would take user input into account.

II. RELATED WORK

The purpose of this study [1] is to use audio and text messages to identify items and educate people. Picture is transformed into text by use of LSTM and sound by means of GTTS. Those who are blind or visually handicapped may use the app to better grasp what's in pictures. Conventional methods of picture captioning are not very good at making broad strokes statements. Because of its superiority and more generalizability, deep learning techniques have

largely replaced more conventional methods. A Convolutional Neural Network (CNN) is used to extract picture characteristics in the suggested approach. Long Short-Term Memory (LSTM) is employed to provide an image description, and the GTTS API is employed to produce audio. Based on its training on the MSCOCO dataset, VGG16 serves as the baseline model. As an encoder-decoder approach, the paper[6] makes use of CNN and LSTM. But in order to glean important details from the caption, they give certain terms varying weights while training the model. On the MS COCO dataset, the suggested R-LSTM model outperformed the state-of-the-art models. In a caption, a heavier weight denotes a more important term. Classifying words according to their topic, status, surroundings, etc., might be facilitated by this. For encoding, we utilize Deep VGG16, and for decoding, we use LSTM.
The article[7] delves into a technique for producing domain-specific captions by using an attention mechanism that takes into account both the object and Issue 07 of Volume 02, 2021 | ISSN: 2582-6832 | Published by the United International Journal of Research and Technology (UIJRT) Information about six attributes is owned by UIJRT.COM. All rights are reserved. There are two components to the suggested methodology: the caption generator and the caption reconstructor. Modified For faster object identification and attribute prediction, RCNN is used. Additionally, caption creation makes use of VGG19 and two stacked LSTMs. One example of a training dataset is MSCOCO. Words that are peculiar to a domain are substituted with generic ones in order to rebuild captions. We have sent the Protege dataset to a semantic ontology expert. The paper's primary method relies on retrieval problems with dense captioning and scene graph matching utilizing structured language descriptions [11]. The most significant issue with caption recommendation systems is the large discrepancy between human perception and information retrieved manually. The Visual Genome dataset was used, and 100 photographs were tagged by hand based on people's perceptions of the images. The language model used is LSTM, while the CNN utilized is VGG16. After that, the dataset is queried using images. An attentional encoder-decoder model that summarizes the news text according to the query picture is proposed in the study [12] as a way for captioning news images. News image captions are different from regular image captions in that they also need to provide information about the picture's backdrop. Using DailyMail corpora, the suggested approach incorporates a simultaneous multi-modal attention mechanism. A gated recurrent unit (GRU) serves as the decoder, while a bidirectional RNN trained using VGGNet serves as the encoder. Facebook and other social media platforms employ the EdgeRank Algorithm[10] to determine the feed's ranking based on how likely people are to like the Starting posts. In response to signals sent by you, the News Feed algorithm takes into consideration factors such as: How often you engage with the friend, Page, or public person (such as an actor or journalist) who published the post The quantity of likes, shares, and comments a post gets from both the general public and your friends specifically2. The amount of time you have spent interacting with this kind of post before. 3. Whether you and other Facebook users are concealing or reporting a certain post. The three main components that make up an Edge are their total. The three of them are Affinity, Decay, and Weight. Your content's EdgeRank and the number of people who view it are directly proportional to the strength of these metrics. Affinity: this metric indicates the degree to which the viewer identifies with the Edge maker. A user's affinity will grow in proportion to the amount of engagement they have with a page's Edges (likes, comments, etc.).A weight is assigned to each kind of Edge, such as a picture, a status update, or a question. We will get into the possibility that publishing more substantial material may raise the EdgeRank later on. Decay is a factor that depends on the age of the edge. As a general rule, EdgeRank is less affected by Edges that are older since they are less valuable. Adaptive Sort[16] primarily reviews existing approaches, such as Genetic Algorithms, and then proposes one to sort huge data sets using Machine Learning. Using ML, we were able to program an algorithm to categorize datasets according to their properties.Our data set's features, comprised of size and pre-sortedness, will feed into our supervised classification learning method.Since they employ Parallel Merge, the Hybrid Approach of Adaptive Sort and EdgeRank Algorithm is the best since it produces the highest number of winning instances. Relevance is one of the factors and parameters.Preciseness Emotionality Comments or suggestions As a parameter, it may be used to acquire better input, which can then be shown in decreasing order.The input of users may also inform the determination of other factors.

III. OUR MODEL

Using deep learning for picture captioning is a part of our approach. We have covered the RNN-LSTM method for caption synthesis and the Faster R-CNN method for object identification. Our suggested system's unique selling point is its ability to take user input into account while generating captions. A picture of a dog taken from[18].

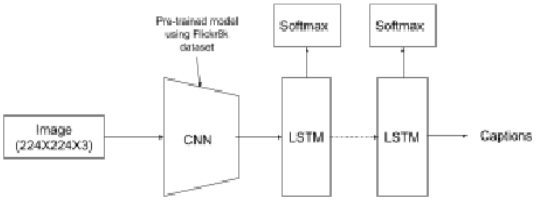

Figure: 1 Dog Image

1) Objection detection using CNN:

To identify many linked things in one frame, to identify multiple associated objects in one frame, object detection is the best approach. Objectn localization is facilitated by it. Various things of interest might be represented by bounding boxes generated around the picture. Convolutional neural networks (CNNs) are a kind of deep learning algorithm that may discriminate based on the weights assigned to certain picture features after receiving an input image. The area CNN is known as R-CNN. The R-CNN method involves breaking the picture into many boxes and then checking each one to see whether it contains an image. Using an input image, the R-Convolutional Neural Network (CNN) model finds the labels and bounding boxes of each item in the picture. When looking for area suggestions, both R- CNN and Fast R- CNN employ selective search. The whole network's performance is negatively impacted by the sluggish and time-consuming process of selective search. Instead of using selective search, Faster R-CNN uses Region Proposal Networks to speed up this procedure. Accordingly, picture captioning tasks requiring object recognition will be best handled by Faster R- CNN. After object classification is complete, the bounding boxes and categorized items will be sent on to the next module.

2)Emotional analysis of images:

This module's primary goal is to generate sentiment-based keywords using convolutional neural networks and to conduct sentiment analysis on user comments in order to enhance the product. When it comes to detecting attitudes, CNN is a top choice. For many computer vision applications, the powerful ResNet model is used. ResNet incorporates the output of one layer into a subsequent layer by use of skip connection. This is useful for resolving the issue of disappearing gradients. First, a big dataset, such as ImageNet, is used to train the CNN. Sentiment prediction model receives the discovered parameters of the pretrained layers and uses them to generate domain-specific picture representations via fine-tuning. Additionally, the CNN method is reliant on the machine learning-based TensorFlow data models. Users' sentimental interpretation of feedback remarks might be categorized as either good, neutral, or negative. Thirdly, RNN-LSTM for caption generation: Long Short Term Memory (LSTM) RNNs are unique among RNNs since they have a memory cell that keeps the data preserved for a longer duration. Because of its internal memory, LSTM is able to process sequential data, take into account both new and old inputs, and recall previously received inputs. The cell state and its many gates form the backbone of LSTM. When it comes to the cell state, the gates are several neural networks that determine what data may be let in. To get the node's output, the gates address complicated RNN issues using matrix transformations, sigmoid activations, and tanh activations. Next, CNN uses the spatial data in the pictures to build what are known as feature vectors. The RNN architecture takes these vectors and uses a fully connected linear layer to create a series of data or words that characterize the picture and create its caption. The goal of the captioning model is to receive an image as input and provide a textual description of it. After passing the input picture through a convolutional neural network (CNN), the RNN will take the output from the CNN and use it to create a descriptive text.

S.N.	Publication	Paper title	Technique	Advantages	Limitations
1	IEEE, 2020	Domain-Specific Image Caption Generator with Semantic Ontology	Modified RCNN + VGG19 + two stacked LSTM + semantic ontology	More relevant captions are generated.	The model is not end-to end in terms of semantic ontology.
2	IJAST, 2020	Image Caption Generator Using Deep Learning	VGG16 + LSTM + GTTS	Higher BLEU scores. Constantly improving by fine tuning the hyper parameter.	Can't predict words out of its library. Less accurate on small dataset.
3	IEEE, 2019	News Image Captioning Based On Text Summarization Using Image As Query	VGGNet + GRU+multi-modal attention mechanism	The multi- modal attention mechanism attends text and image simultaneously. GRU is less time consuming, uses less memory.	GRU lacks accuracy when datasets have longer sequences.
4	ScienceDirect, 2019	Liking, sharing, commenting and reacting on Facebook: User behaviors' impact on sentiment intensity	EdgeRank Algorithm	Various Parameters can be Considered	Might be sometimes be false depending on decay rate.
5	IEEE, 2017	Image Retrieval By Dense Caption Reasoning	VGG16 + LSTM + scene graph matching	The model achieves better performance and is more effective. The captions generated are more relevant.	The task of manually labeling images can be tedious for larger datasets.
6	AAAI, 2017	Reference Based LSTM for Image Captioning	VGG16 + R-LSTM + k-Nearest Neighbor	Outperforms existing techniques with better accuracy.	Relying too much on references can lead to poor performance. Assigning weights makes it a little complex
7	IJCSIT, 2016	Adaptive Sorting Using Machine Learning	Parallel Merge Sort	Best for overall characteristics i.e size and pre-sortedness.	Require Multithreading and Multiprocessing environment.

An picture must be selected before a description can be generated.data inputted into a convolutional

neural network (CNN) architecture, such as ResNet. At a softmax classifier is located at the network's terminus and a vector of class scores is produced as a result. But there's only a collection of characteristics that convey the location-based information in the picture is necessary. In order to acquire spatially-oriented material, the the last fully-connected layer that does the picture classification
should be eliminated and the results from the prior layer used and efficiently utilizes geographical information from which the RNN-LSTM model is fed. Fourthly, Adaptive Sorting for

Recommendation of Captions:

Showing off the top AND most relevant captions for the user. The descriptionproduced by the prior module must to be organized in rank in descending order of precision. The description determined by examination of the past actions taken by the user, often the kind of based on his past records, he generally like certain captions. and the results of sentiment analysis analysis. Extra parameters for the Edge-Rank Algorithm on what is important to think about while organizing and Utilize parallel merge sort. Prior to
Criteria might be based on comments and sentiment analysis. in terms of edge selection and criteria such as correctness, relevancy, opinions and comments may be used to allocate advantages weights. Deterioration is related to the frequency of app use. as well as the amount of time spent using the program. continues to assess the user's preferences. The When ranking captions, feedback is also considered. It greatly enhances the likelihood of obtaining top

the first page's captions.

IV.Conclusion

The material may be enhanced by adding written descriptions with photographs - foundational picture retrieval effectiveness, the growing use case for visual comprehension in the domains of health, safety, military, and theoretical framework in addition to the procedures for picture captions determined by a number of criteria, including feedback, relevance, precision, etc., by using adaptive Method for sorting. Additionally, the last input continues continually studying the user's preferences and preferences of the individual using it. Additionally, the comments are considered remember to arrange captions, which greatly enhances the likelihood of receiving top captions on the first page.

REFERENCES

Tehseen Zia, Shahan Arif, Shakeeb Murtaz,Mirza Ahsan Ullah. “Text to image generation with attention based recurrent neural networks.” arXiv: 2001.06658,2020.
[2] Niange Yu, Student member, IEEE , Xiaolin Hu ,Senior Member , IEEE , Binheng Song , Jian Yang , and Jianwei Zhang. “Topic Oriented Image Captioning Based on Order Embedding, Image processing”. Volume.28, no-6 , JUNE 2019.
[3] Chetan Amritkar and Vaishali Jabade. “Image Caption Generation using Deep Learning Technique”. 25th April 2019.

UIJRT | United International Journal for Research & Technology | Volume 02, Issue 07, 2021 | ISSN: 2582-6832
All rights are reserved by UIJRT.COM. 9
[4] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. “An Overview on Image Sentiment Analysis: Methods, Datasets and Current Challenges”. 2019.
[5] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz
Shiratuddin and Hamid Laga. “A Comprehensive Survey of Deep Learning for Image Captioning. arXiv: 1810.04020v2, 14th October, 2018.”
[6] Marco Pedersoli, Thomas Lucas, Cordelia Schmid
and Jakob Verbeek. “Areas of Attention for Image Captioning”. 2017

[7] Jyoti Islam and Yanqing Zhang. “Visual Sentiment
Analysis for Social Images Using Transfer Learning Approach”. October, 2016.
[8] Aaron van den Oord, Nal Kalchbrenner, and Koray
Kavukcuoglu. Pixel recurrent neural networks. arXiv preprint captioning can cause the development of the theory an arXiv:1601.06759, 2016
[9] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo
Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623, 2015
[10] Stuti Jindal and Sanjay Sin10.1051/matecconf/201820000020
[11] Ruslan Salakhutdinov. Learning deep generative
models. Annual Review of Statistics and Its Application, 2: 361–385, 2015.
[12] Alec Radford, Luke Metz, and Soumith Chintala.
Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.
[13] Vasavi Gajarla and Aditi Gupta. “Emotion Detection and Sentiment Analysis of Images”. 2015
[14] Gregor Blossey, Jannick Eisenhardt, Gerd Hahn,
“Blockchain Technology in Supply Chain

Management:An Application Perspective”, doi:10.24251/HICSS.2019.824