Caption Recommendation System

Jayesh Asawa[1], Mansi Deshpande[2], Sampada Gaikwad[3] and Riddhi Toshniwal[4]

SCTR's Pune Institute of Computer Technology (PICT), Pune, Maharashtra, India

Email: [1]jayeshasawa1@gmail.com, [2]mansideshpande27@gmail.com, [3]gaikwadsampada8@gmail.com and [4]riddh.toshniwal@gmail.com

Abstract— Caption generation is the challenging neural network problem of generating a human-readable textual description to the given photograph. It requires understanding from the domain of computer vision as well as from the field of natural language processing. Every day, we encounter a large number of images on social media. These sources contain images that viewers would have to interpret themselves. Image captioning is important for many reasons. For example, Facebook and Twitter can directly generate descriptions based on images. The descriptions can include what we wear, where we are (e.g., beach, cafe), and what we are doing there. To generate automatic captions, image understanding is important to detect and recognize objects. It also needs to understand object properties and their interactions with other objects and scene type or location. Generating well-formed sentences requires both semantic and syntactic understanding of the language. In deep learning based techniques, features are learned automatically from training data and they can handle a large set of images and videos. Deep learning techniques such as CNN will be used for image classification and RNN encoders and decoders will be used for text generation that is captions for the provided image. Language models such as LSTM will also be implemented in both sentiment analysis and caption generation.

Keywords— Computer Vision, Automatic captions, Semantic, SyntacticDeep learning, CNN, RNN, LSTM, Sentiment analysis, Caption generation.

I. INTRODUCTION

As there are humongous numbers of social media users, our project focuses on generating appropriate captions for the image(s) provided by the user. For example, platforms like Facebook and Instagram can infer directly from the image, where you are (location: beach, cafe etc), what you wear (color) and more importantly what you're doing also. The ability of advanced technologies like Artificial Intelligence and Machine Learning to leverage the information provided by human beings for in depth analysis and insights of given images, including prediction of the meaningful captions is very interesting and motivating. Apart from this image captioning has interesting applications like Picasa, Tesla (Google self drive cars), SkinVision (Is used to confirm whether a skin condition can be skin cancer or not.) and Google photos(Classify your photo into different categories like Mountains, sea etc.) etc. It can also serve as a huge help for visually impaired people. Proposed methodology includes use of deep learning for image captioning. In contrast to hand crafted descriptors used in traditional detectors, deep convolutional neural networks generate hierarchical feature representations from raw pixels to high level semantic information, which is learned automatically from the training data and shows more discriminative expression capability in complex contexts. For object detection, we will be using Faster R-CNN technique and for caption generation, RNN-LSTM technique has been discussed. Our proposed system will be consideration of feedback for caption generation.

II. RELATED WORK

This paper [1] aims to identify objects and inform people with the help of audio and text messages. Image is converted to text using LSTM and audio using GTTS. The application is built for blind people and the visually impaired to help them understand images. The traditional approaches of image captioning struggle to generalize. Deep learning techniques prove better than the traditional approaches and are more generalized compared to traditional approaches. In the proposed methodology, Convolutional Neural Network to extract features from the image, LSTM is used to generate description of the image and GTTS api is used to generate audio. The baseline model used is VGG16 which is trained on the MSCOCO dataset.

The paper[6] uses CNN and LSTM encoder- decoder technique. However, while training the model, they assign different weights to different words to extract key information from the caption. The proposed R- LSTM model has been tested on MS COCO dataset and is better than existing models. A larger weight indicates higher significance of the word in the caption. This can help classify words as the main subject, its status, its environment, etc. Deep VGG16 is used as an encoder and LSTM is used as a decoder.

The paper[7] discusses a method of generating domain specific captions which generate captions using attention mechanism which considers both object and attribute information. The proposed methodology consists of two parts: Caption generator and caption reconstructor. Modified Faster RCNN is used for object detection and attribute prediction. VGG19 and two stacked LSTMs are further used for caption generation. MSCOCO is used as a training dataset. For caption reconstruction, specific words in the general caption are replaced with domain specific words. Protege dataset has been referred for semantic ontology.

The main strategy used in the paper[11] is using dense captioning and scene graph matching issues by using structured language descriptions for retrieval. The main problem while recommending captions is the wide gap between manually extracted features and high- level human perceptions. The dataset used is Visual Genome dataset and 100 images from the dataset are manually labeled according to people's understanding of the image. The CNN used is VGG16 and the language model used is LSTM. Images are then provided as a query to the dataset.

The paper[12] proposes a news image captioning method using an attentional encoder- decoder model which summarizes the news text according to the query image. Captions of news images differ from normal images as they should contain information about the background of the image as well. The proposed methodology includes a simultaneous multi- modal attention mechanism on DailyMail corpora. The encoder is bidirectional RNN using VGGNet and decoder is GRU (gated recurrent unit).

EdgeRank Algorithm[10] used by Facebook and other social networking sites to rank the feed accordingly, so that users are most likely to like the Starting posts. The News Feed algorithm responds to signals from you, including, for example:

- How often you interact with the friend, Page, or public figure (like an actor or journalist) who posted
- The number of likes, shares and comments a post receives from the world at large and from your friends in particular
- How much you have interacted with this type of post in the past
- Whether or not you and other people across Facebook are hiding or reporting a given post. Each Edge is made up of the sum three key factors that form it. These are Affinity, Weight and Decay. The higher each of these factors is, the higher the EdgeRank and the more people will see your content.

Affinity – this is a measure of how 'close' the viewing user is to the Edge creator. If a user makes a number of interactions (likes, comments etc) with the Edges of a particular page then their affinity will be greater.

Weight – each type of Edge (e.g. photo, status update, question) is given a weighting. Publishing content of a heavier weight may increase the EdgeRank, which we'll explore further below. Decay – the factor based on how long ago the edge was created. Broadly speaking, the older the Edge the less value it has, therefore the less impact it has on EdgeRank.

Adaptive Sort[16] is mainly aimed at proposing a method to sort large data sets using Machine Learning while reviewing other methods such as Genetic Algorithms. Machine Learning allows us to build an algorithm to sort data sets based on their characteristics. The characteristics i.e size and pre-sortedness, will be the features of our data set, which acts as the input to our supervised classification learning algorithm.

Hybrid Approach of Adaptive Sort and EdgeRank Algorithm uses Parallel Merge and can be considered best, as the number of winning instances are maximum in them. Factors and Parameter like,

- Relevance
- Accuracy
- Sentiments
- Feedbacks

Can be used as a parameter to get better input and then we can display based on descending order of input. Other parameters can be decided based on user feedback.

III. OUR MODEL

Our model includes use of deep learning for image captioning. For object detection, we will be using Faster R- CNN technique and for caption generation, RNN-LSTM technique has been discussed. The special feature of our proposed system will be consideration of feedback for caption generation. Dog Image from[18].

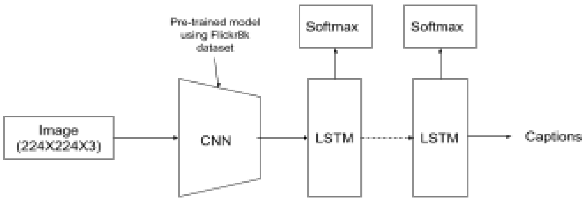

Figure: 1 Dog Image

1) Objection detection using CNN:

To identify multiple related objects in one frame, to identify multiple related objects in one frame, object detection is the best technique. It provides localization of objects. Many bounding boxes could be formed around the image representing different objects of interest. CNN (Convolutional neural networks) is a deep learning algorithm which can take in an input image, assign appropriate weights to particular aspects of the image and differentiate on the basis of it. R- CNN is region CNN. In R- CNN, the image is divided into a bunch of boxes and the boxes are checked to find out if they contain an image or not. R- CNN model takes and image as input and identifies bounding boxes and labels for each object in the image. Both of the above algorithms (R- CNN & Fast R- CNN) use selective search to find out the region proposals. Selective search is a very slow and time-consuming process affecting the performance of the entire network. Faster R- CNN speeds up this process by making use of Region

Proposal Networks instead of selective search. Hence, Faster R- CNN will be most suited for object detection in image captioning. Finally, the objects classified along with bounding boxes will be sent to the next module for further processing.

2) Sentimental analysis of image:

Generation of keywords using sentiments by using convolutional neural networks and sentimental analysis of feedback given by user to improve the product is the main objective of this module. CNN is one of the best ways to detect sentiments. Residual Neural Network (ResNet) is a powerful model which is used in many computer vision tasks. ResNet makes use of skip connection to add the output from an earlier layer to a later layer. This helps to solve the vanishing gradient problem. For this, the CNN is first trained on some large dataset like ImageNet. The identified parameters of the pretrained layers are transferred to the sentiment prediction model for generating image representations using domain specific fine tuning. The CNN technique is also dependent on TensorFlow data models that are based on machine learning technology. Sentimental analysis of feedback statements given by users can then be classified as negative, positive and neutral response.

3) RNN-LSTM for generating captions:

LSTM (Long Short Term Memory) is a special kind of RNN that includes a memory cell which ensures that the information is maintained for a longer period of time. LSTM can handle sequential data, it considers current input as well as previously received inputs, and can memorize previous inputs due to its internal memory. The core of LSTM lies in the cell state, and its various gates. The gates are different neural networks that decide which information should be allowed on the cell state. The gates perform matrix transformations, sigmoid and tanh activations in order to solve complex RNN problems and get the output of the node. Vectors called feature vectors are then generated from the spatial data in the images by CNN. These vectors are then fed through a fully connected linear layer into the RNN architecture in order to generate sequential data or sequence of words that describe the image and formulate its caption. The captioning model should take in an image as input and output a text description of that image. The input image, processed by a CNN will then connect the output of the CNN to the input of the RNN and generate 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.

In order to generate a description, a particular image is fed into a pre-trained CNN like ResNet architecture. At the end of this network, a softmax classifier is present and it outputs a vector of class scores. However, only a set of features that represents the spatial content in the image is required. To get that kind of spatial content, the final fully connected layer that classifies the image should be removed and the output of the previous layer which distills spatial information can be used effectively for feeding it to the RNN- LSTM model.

4) Adaptive Sorting for Caption Recommendation:

This module mainly focuses on displaying the best and most relevant captions to the user. The captions generated by the previous module should be sorted in decreasing order of accuracy. The captions recommended will be displayed by analysing the previous behaviour of the user, generally what type of captions he tends to like based on his previous records and feedback obtained by performing sentiment analysis. Edge-Rank Algorithm with extra parameters about relevance should be considered for sorting and parallel merge sort should be used. The previous feedback and sentiment analysis can be used as criterion for edge selection and factors like relevance, accuracy, sentiments and feedback can be used to assign edge weights. For decay, the number of times the app is used and time spent on the app can be considered.

keeps analysing the likes, dislikes of the user. The feedback is also taken into account for caption sorting which definitely improves the chances of getting best captions on the first page.

IV. CONCLUSION

The images text description can improve the content-base image retrieval efficiency, the expanding application scope o visual understanding in the fields of medicine, security, militar and also, theoretical framework and also the methods of image captions based on various factors like relevancy, feedbacks, accuracy etc using EdgeRank Algorithms and Adaptive Sort approach. Also, the feedback at the end keeps on updating the taste of the user, and keeps analysing the likes, dislikes of the user. The feedback is also taken into account for caption sorting which definitely improves the chances of getting best captions on the first page.

REFERENCES

[1] 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.

[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