Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 7
July 2023

Image Captioning Using Deep Learning

[1] S.T.Santhanalakshmi, [2] Dr. Rashmita Khilar
[1] Assistant Professor, Department of Computer Science and Engineering, Panimalar Engineering College, Chennai,Tamilnadu , India
[2] Associate Professor, Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of
Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India
anandh.shantha@gmail.com
rashmitakhilar.sse@saveetha.com

Abstract: People today need captions for a variety of purposes, including sharing a picture to social media, creating news headlines based on an image, among other things. Instead of manually creating captions for each image, an image captioning system aims to generate them automatically. It provides a descriptive statement for an image that aids in the semantic interpretation of the visual. Picture understanding is a crucial method for decoding semantic image data, and VGG16 can use it. Natural language processing and computer vision are combined in the process of image captioning. Either a standard machine learning approach or a deep learning strategy can be used to achieve the goal. Identifying items and determining the relationships between them are essential for carrying out the planned job. Feature extraction is a technique for converting the image into a vector for further processing. The LSTM receives the items and visual material and connects the words to create a sentence that describes the objects. The implementation approach for Object Detection-based Picture Captioning using Deep Learning is presented in this work. we have employed the CIDEr metrics while evaluating; the accuracy of 94.8% is achieved for 30 epochs which is significantly good. For image classification, CNN and the Flickr dataset are utilized.

Keywords*: Object Detection, Deep Learning, Computer Vision, Image Captioning*.

1. Introduction

The ultimate aim of an image captioning assignment is to use sentences to depict an image. Natural language processing and computer vision are both necessary for this.[1-3] Knowing the semantics of the objects and other visual information, as well as having understanding of natural language processing, is the most difficult challenge in the process of generating captions. The creation of a relationship between the extracted objects is necessary for sentence production. [4-5]The process of extracting features can be used to extract image data. An image comprises a variety of information, including the objects in the picture and their significance. The two main categories of image understanding are deep learning techniques and traditional

machine learning approaches. With conventional machine learning methods, the Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradient can be used to extract features (HOG).[6-8] Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are employed in deep learning approaches to generate captions.

We used the YOLO (You Only Look Once) method for object detection in this study to demonstrate how to perform image captioning.[9-10] The bounding box and confidence for an object are given when it is discovered. For an image to be properly understood, image data is just as crucial as the objects themselves. Features extraction is the process of removing important traits and important picture data from an image. The feature extraction process employs the VGG16 and InceptionV3 types of convolutional neural networks (CNNs). The object and picture characteristics created by these methods are fed into the Recurrent Neural Network (RNN), which uses them to form sentences. They are the input for RNN, which outputs a variable-length description. A textual description of a picture is produced by RNN.The humble beginnings of captioning for images could be found in a number of practical applications, such as the reality that it is capable of converting text descriptions into audio for those who are blind or suffer from poor eyesight and rely on audio. It is also used for video subtitles, self-driving cars, effective image search while surfing, and deciphering newspaper articles. The biggest challenge in creating a picture description starts with object recognition in the image applying statistical language models and static object class libraries.

CNN being used: It's a Deep Learning system that will take a 2D matrix input image, give distinct aspects and objects in the image weights and biases that can be learned, and be smart enough to discern one from the other. This approach was helpful in identifying the objects in an image, but it was unable to tell us how those objects related to one another (that is just basic image classification). Here, an in-depth recurrent infrastructure-based generative model that includes recent developments in visual analysis and automatic translation is used to produce meaningful sentences. bigUsing an RNN: These networks have loops in them, which enable the persistence of information. An RNN variant that can recognize long-term dependencies is the LSTM.

2. Related Work

In their study, [11] suggested a region-based method for captioning pictures. The region-based objects that are used for object detection implement the image captioning (ROD). The region-based technique divides the picture into areas for object detection. Authors specifically employ Region Based CNN, also known as R-CNN. Other CNN-based techniques used in the study work include scene classification and feature extraction. With the aid of object, scene, and image feature properties, the RNN approach is employed to generate sentences. Instead of using a straightforward RNN, [12] suggested a model that use LSTM. Feature Extraction was employed in the

investigation using the pre-trained VGG16 model. LSTM can be used in place of RNN since LSTM can recall the words for a longer amount of time than RNN can. A cell called the LSTM holds the words until the appropriate caption can be created. This is supported by the findings in the categories of captions that are related to one another and those that are not related. These classifications result from taking the surrounding words into account. The Show and Tell approach for picture captioning was proposed by [13]. The Show and Tell technique employs both neural machine translation and picture recognition. The Inception-v3 model and LSTM are combined in the suggested approach. During the captioning process, the LSTM cell is used to hold auxiliary words, while the Inception-v3 model handles object recognition. The Inception-v3 model processes the input image for object recognition first, creating the vector representation of the image. This vector is utilized by the LSTM to produce descriptions. A Word Level Attention model for image captioning was put up by [14]. In order to process picture information with two models for precise word prediction, the model contains a word-level attention layer. The two modes are the attention mode, which extracts word-level attention, and the line level bidirectional spatial encoding, which is used for feature maps.. Additionally, it made use of bidirectional LSTM networks, which process words in both forward and backward orientations. The Word Level Attention Extraction approach harvests visual information to forecast the next word using a softmax activation function. Better phrase learning using Deep Convolutional Network is the emphasis of [15] study. The method for extracting picture features is a deep Fisher Kernel one. A Fisher Vector is created by aggregating the retrieved activations. This method makes use of gLSTM in place of LSTM. As the process progresses, less is known about the input image because information about it is only supplied at the first step. The gLSTM provides an additional piece of information known as a "guide" that is related to the input picture data throughout the process. [16-20] suggested a method that allows the usage of Read-Only LSTM with LSTM. The LSTM cell is a storage device that enables the storing of intermediate words, whereas the read-only LSTM cell provides image features. The current word must be forecasted in relation to the visual content, which is only provided in the first phase, using a separate unit that associates both the previous and current words. The LSTM with Read-Only Unit, a new extra unit designed, increases accuracy.

The NIC(Neural Image Caption) model was applied by Vinyals, Oriol, et al. The encoder used in the NIC model is CNN. [21-24]The pre-trained CNN's final layer is fed into the RNN decoder, which uses the network to categorize the images. This RNN decoder creates more sentences. They made use of LSTM, a more sophisticated RNN. Recently Xuetal suggested adding visual attention summarization to the LSTM model to enable it to concentrate its gaze on distinct objects while generating related

words. For creating captions for images that are human-like, neural language models are helpful. With the exception of the newest methods, they all adhere to a similar encoding-decoding structure that combines caption creation with visual attention.[25] This work dealt with the third category of caption creation techniques. In our work we have designed a model which produces explanations for images in natural language. The image encoder CNN is employed. The RNN decoder generates the phrase using the last hidden layer as input prior to pre-training for image classification tasks.[26] This research presents a neural architecture for caption generation from images. The foundation of this system is mostly taken from probability theory. It is feasible to get better results by employing a robust mathematical model, which maximizes the likelihood of the right translation for both inference and training. For our model to work, we need a dataset that includes both the captions and the images. The dataset should allow the training of the picture captioning model. [27]The Flickr8k dataset is a benchmark dataset that is openly available for picture to sentence description. This collection consists of 8000 images, each with five captions.

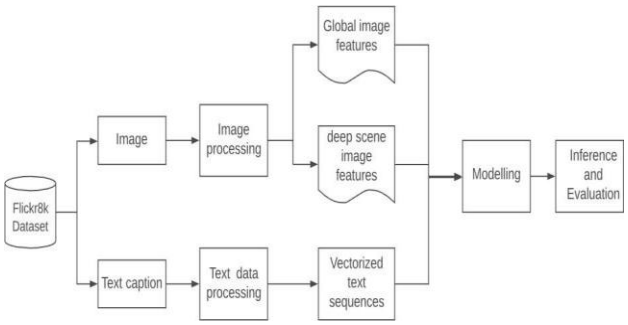


Fig.1 Working flow of the model

3. Image Data Preparation

The Flickr website's images come from a variety of groups. The things and behaviors depicted in each caption are clearly described. The lack of photographs of famous individuals and locations makes the dataset more generic as it shows a diversity of events and circumstances. 1000 photos are in the development dataset, 1000 images are in the test dataset, and 6000 images make up the training dataset. The dataset has the following characteristics that make it appropriate for our job: The model becomes generic and avoids getting over-fitted when multiple captions are mapped to a single image. The image captioning model can be made more robust by using a variety of training picture types that cover a range of image types.

In order to train a deep learning model, the image should be transformed into relevant features. Before any image is taught using a deep learning model, features must first be extracted from the image. A convolutional neural network (CNN) with the Visual Geometry Group (VGG-16) model is used to extract the

features. This model also won the 2015 Image Net Large Scale Visual Recognition Competition by successfully classifying the photographs into one of the 1000 classes offered in the challenge. Therefore, since image captioning calls for the identification of images, this model is usage of 3*3 convolutional layers, and it employs a maximum pooling layer between each layer to lower the image's volume size of 24. After the last layer of the image that predicts categorization has extracting properties from the input image, which must be 224x224. For each photograph in the Flickr8k collection, numerous descriptions are provided. Each perfect for usage in this project. The deeper number of layers in the VGG-16 network, which has 16 weight layers, aids in more accurate extracting features from photos. The VGG-16 network's design is straightforward due to the

been eliminated, The feature is the internal representation of the picture just prior to categorization. This model creates a 1-dimensional 4096 element vector by

image's id is used as a key during the data preparation stage, and a dictionary is used to store the captions that go with each one as values.

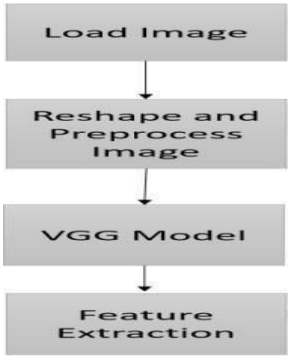


Fig.2 Caption Data Preparation

Raw text must be converted into a format that machine learning or deep learning models can understand. The text is cleaned as follows before to use. Eliminating punctuation deleting the numbers, eliminating words of a single length characters are changed from uppercase to lowercase. Stop words are included in the text data because removing them would make it more difficult to create a caption that is grammatically correct.

Original Captions	Captions after Data cleaning
Two people are at the edge of a lake, facing the water and the city skyline.	two people are at the edge of lake facing the water and the city skyline
A little girl rides in a child 's swing.	little girl rides in child swing
Two boys posing in blue shirts and khaki shorts.	two boys posing in blue shirts and khaki shorts

Table 1. Data cleaning of captions

For picture captioning, there are three databases with a wide variety of image formats. Eight thousand images make up the Flickr8K Dataset, of which six thousand were utilised for training, one thousand for validation, and one thousand for testing. The Flickr30K Dataset consists of 30K total photos, 28K for training, 1K for validation, and 1K for testing. Microsoft unveiled MSCOCO, a dataset for object identification and picture captioning. 82783 training images, 40504 validation images, and 40775 testing images are included in the 328K total number of images that make up MSCOCO. Experimental results in preprocessing showed that classification with grey scale images resulted in higher accuracy classification than with RGB images. Its helps in simplifying algorithms and as well eliminate the complexities related to computational requirements. The noise reduction is done with Gaussian filter which is a low pass filter passed through each pixel to find the edge of the object that remains clear. Our Model used Morphological technique Dilation to make objects more visible and fills in small holes in object and hence making lines thicker and filled shapes appear larger. By determining how similar each word is to a reference sentence, the resulting sentences are assessed using various metrics.

4. Architecture Overview

This shows the entire layout of our working model, including all of its parts and the states that exist when the process is being executed. The very first step of the picture feeding process is shown, followed by the parsing and conversion of the image into vectors. All of the image-related information is recorded and fed to the model using the Flickr dataset. In order to construct the caption with the aid of language processing and data that has been trained and saved, CNN is used for the encoding and STM for the decoding of the descriptive data that play the image again while ageing. This results in the output of a generated caption.

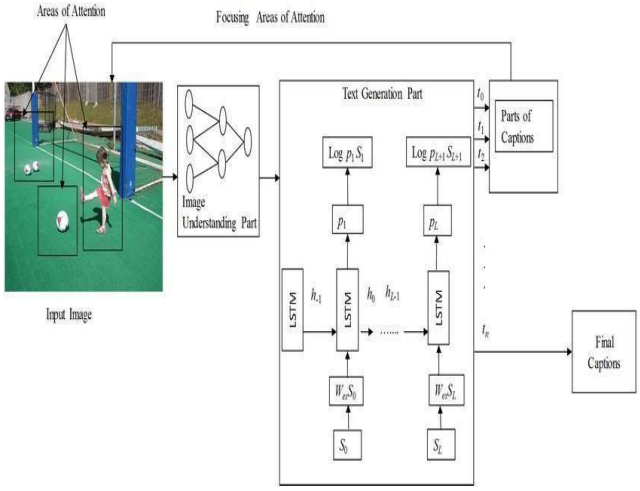


Fig.3 Architecture overview

Data preprocessing - images:

The only thing that images are to our model is input (X). Any input to a model must be provided as a vector, as you may already be aware. Every picture must be neural networks.

This model was trained using the Visual Genome Corpus dataset to perform image classification on 1000 different kinds of pictures. However, our goal is to just obtain a fixed-length informative vector for each image, not to categorize the image. The term "automated feature engineering" refers to this procedure.

Data preprocessing - captions:

Here, the challenge is predicting the captions. As a consequence, during the training phase, captions will serve as the target variables (Y) that the model is learning to forecast. The entire caption, however, cannot be foreseen based solely on the image. Each word in the caption is predicted. Therefore, each word must be encoded into a fixed-sized vector.

And so on… The main text file which contains all image captions is converted into a fixed-size vector before the neural network can utilise it as input. We choose transfer learning for this purpose utilizing Google Research's Inception V3 model of convolutional

Data preparation using generator function:

Take the first picture, Image 1, which has the caption "start seq the black cat sat on grass end seq" as its associated text. Keep in mind that the input is the picture vector, and the prediction is the title. However, the caption that follows is expected:

First, the picture vector and the first word are provided as input, and then the second word is attempted to be predicted, i.e.Input=Image_1+‘startseq’;

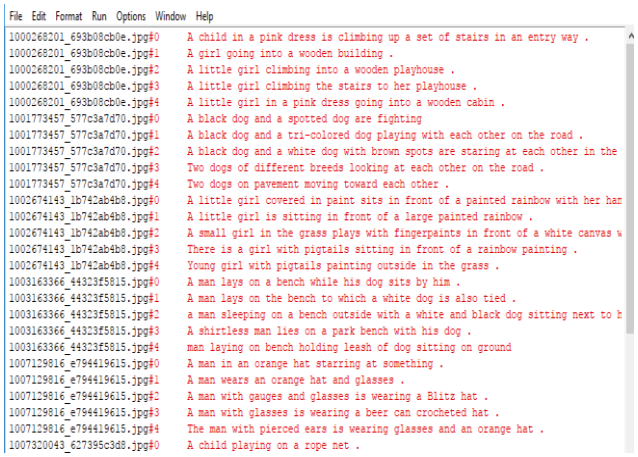
Output=‘the’

Following that, an image vector and the first two words are given as input, and the third word is attempted to be predicted, i.e.

Input = Image_1 + ‘startseq the’;

Output = ‘cat’

Flickr8k.token in our **Flickr_8k_text** folder.



```
File  Edit  Format  Run  Options  Window  Help
1000268201_693b08cb0e.jpg#0     A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1     A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2     A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3     A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4     A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg#0     A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg#1     A black dog and a tri-colored dog playing with each other on the road .
1001773457_577c3a7d70.jpg#2     A black dog and a white dog with brown spots are staring at each other in the
1001773457_577c3a7d70.jpg#3     Two dogs of different breeds looking at each other on the road .
1001773457_577c3a7d70.jpg#4     Two dogs on pavement moving toward each other .
1002674143_1b742ab4b8.jpg#0     A little girl covered in paint sits in front of a painted rainbow with her han
1002674143_1b742ab4b8.jpg#1     A little girl is sitting in front of a large painted rainbow .
1002674143_1b742ab4b8.jpg#2     A small girl in the grass plays with fingerpaints in front of a white canvas w
1002674143_1b742ab4b8.jpg#3     There is a girl with pigtails sitting in front of a rainbow painting .
1002674143_1b742ab4b8.jpg#4     Young girl with pigtails painting outside in the grass .
1003163366_44323f5815.jpg#0     A man lays on a bench while his dog sits by him .
1003163366_44323f5815.jpg#1     A man lays on the bench to which a white dog is also tied .
1003163366_44323f5815.jpg#2     a man sleeping on a bench outside with a white and black dog sitting next to h
1003163366_44323f5815.jpg#3     A shirtless man lies on a park bench with his dog .
1003163366_44323f5815.jpg#4     man laying on bench holding leash of dog sitting on ground
1007129816_e794419615.jpg#0     A man in an orange hat starring at something .
1007129816_e794419615.jpg#1     A man wears an orange hat and glasses .
1007129816_e794419615.jpg#2     A man with gauges and glasses is wearing a Blitz hat .
1007129816_e794419615.jpg#3     A man with glasses is wearing a beer can crocheted hat .
1007129816_e794419615.jpg#4     The man with pierced ears is wearing glasses and an orange hat .
1007320043_627395c3d8.jpg#0     A child playing on a rope net .
```

Fig.4 Flicker Data project Set text format

Extracting the feature vector from all images:

Using pre-trained models that have already been trained on huge datasets, this method extracts their properties and applies them to our problems. Transfer learning is another name for it. We employ the Xception model, which was created using the imagenet dataset, which consists of 1000 different classes. Direct import of this model is possible from keras applications. Having an internet connection is essential because the weights are downloaded automaticallySince the Xception model was first developed for imagenet. It's important to know that the Xception model takes input for photos up to 299*299*3 in size. 2048 feature vectors are obtained once the last classification layer is removed.Model = Xception (include_top = False, pooling = "avg")

All images' features will be extracted by the function extract features(), which also associates each image's name with a corresponding feature array. The features dictionary is then dropped into a pickle file called "features.p." Depending on your system, this process could take a long time. It will take about 7 minutes to complete this assignment using an Nvidia 1050 GPU for training purposes. This procedure takes 4 to 5 hours because to the CPU. Code may be commented, and features may be loaded straight from a pickle file.

x1 (feature vector)	x2 (Text sequence)	y (word to predict)
feature	start,	two
feature	start,two	dogs
feature	start,two, dogs	drink
feature	start,two,dogs, drink	water
feature	start,two,dogs,drink, water	end

Table 2. Word Prediction Generation Step by Step

Training the model:

6000 training photos are delivered in batches, and the model is then used to fit the data.To produce the input and output sequences, fit the generator() function. After that, save this model to a folder; this process may take some time.

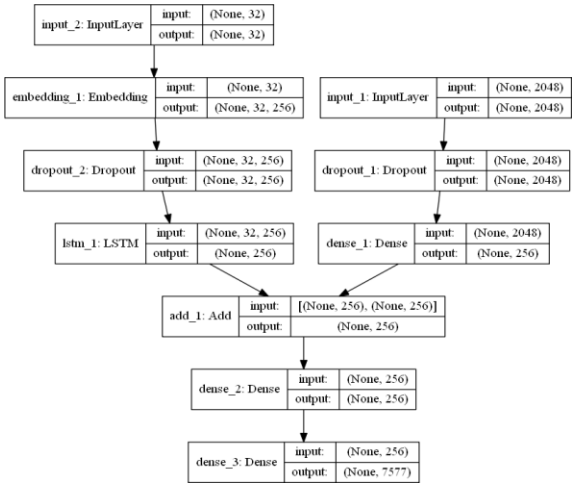


Fig.5 Final Model Structure

5. PERFORMANCE ANALYSIS

BLEU

A quality metric score for MT systems called BLEU aims to gauge how closely a translation produced by a computer and one produced by a human coincide. The fundamental tenet of BLEU is that a machine translation is better the more closely it resembles a qualified human translation. Only the source sentences and translations chosen for the test are used to evaluate a system's performance in BLEU tests. It is frequently feasible to score good translations negatively since the translation that was chosen for each segment may not be the only accurate one. As a result, especially when it comes to information that differs from the specific test material, the ratings don't always accurately reflect a system's prospective performance.

ROUGE

By comparing overlapping n-grams, word sequences, and word pairs, we use ROUGE-L, which effectively measures the longest comm, as it was previously established by Lin (2004) to assess summarization systems. Between each pair of sentences in this sub-sequence, Long phrases are preferred by ROUGE since it heavily relies on recollection, as well highlighted by (Vedantamet al., 2015).

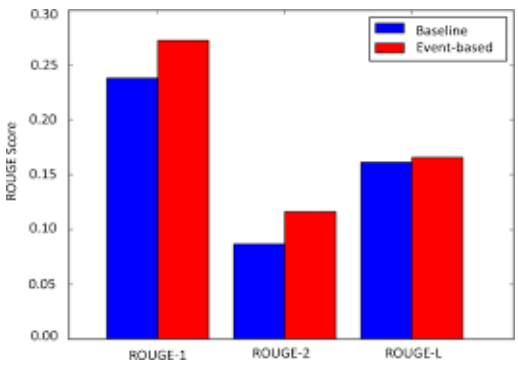


Fig.6 Graph comparing metrics

CIDEr

CIDEr(Consensus - based Image Description Evaluation) (Consensus - based Image Description Evaluation) sentences that were authored by humans. This indicator shows that the majority of people, as determined by the measure, strongly concur.

Accuracy

While analysing the accuracy of 94.8% for 30 epochs, we use the CIDEr metrics. In particular, human participants are provided a triplet of descriptions for (Consensus-based Image Description Assessment) This metric evaluates how much a sentence that was generated resembles a group of ground truth evaluation purposes—one reference and two candidate descriptions—and asked to choose the candidate description that most closely resembles the reference. If a measure awards a greater score to the human subject's selection of the reference caption as being more comparable, it is considered to be correct..

Fig.7 Testing image 1

TESTING

Paragraph: On the road, five horses are running. One by one, the horses are running in a line. There are clouds and blue sky. The horses have brown and white colour. At a considerable distance from the horses, there are poles.

Fig.8 Testing image 2

Paragraph: On the grass, a girl and another person are seen. A person is holding an umbrella while wearing a black shirt and black pants. A girl wearing a yellow dress, white shoes, and a cap. Behind the two people, there is a long tree.

Table 3. Experiment Result By determining how similar each word is to a reference sentence, the resulting sentences are assessed using various metrics.

	Expert Annotations	Crowd Flower
BLEU-1	0.191*	0.206
BLEU-2	0.212	0.212
BLEU-3	0.209	0.204
BLEU-4	0.206*	0.202
METEOR	0.308*	0.242
ROUGE-L	0.218*	0.217
CIDEr	0.289*	0.264

Table 4. Different methods are performed against different evaluation metrics.

SENTENCES	BLEU	CIDER	ROUGE	METEOR
S1	0.579	0.600	0.396	0.195
S2	0.404	0.658	0.274	0.256
S3	0.279	0.599	0.400	0.172
S4	0.191	0.677	0.450	0.137

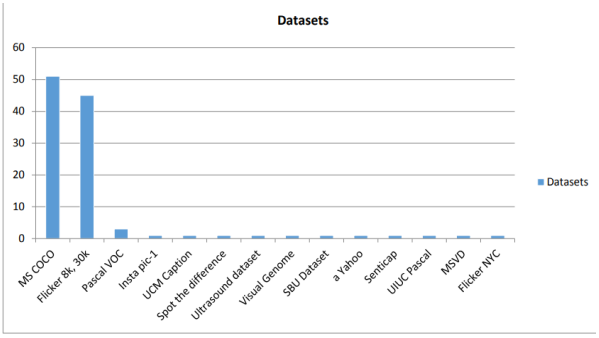


Fig 9 Efficient dataset

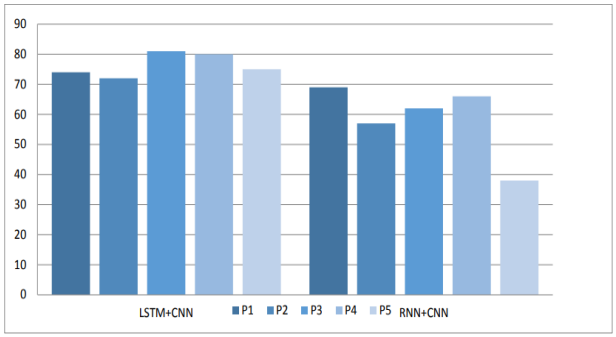


Fig.10 Usage of LSTM+CNN model compared with the usage of RNN+CNN model

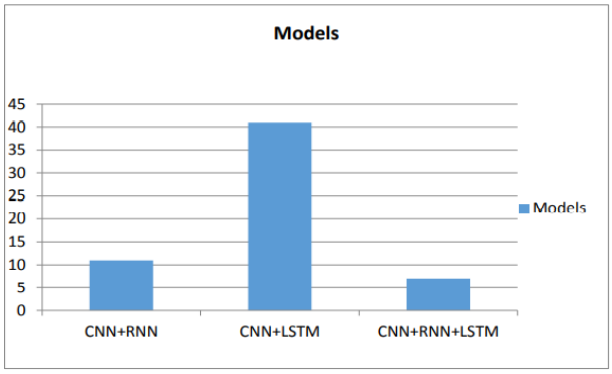
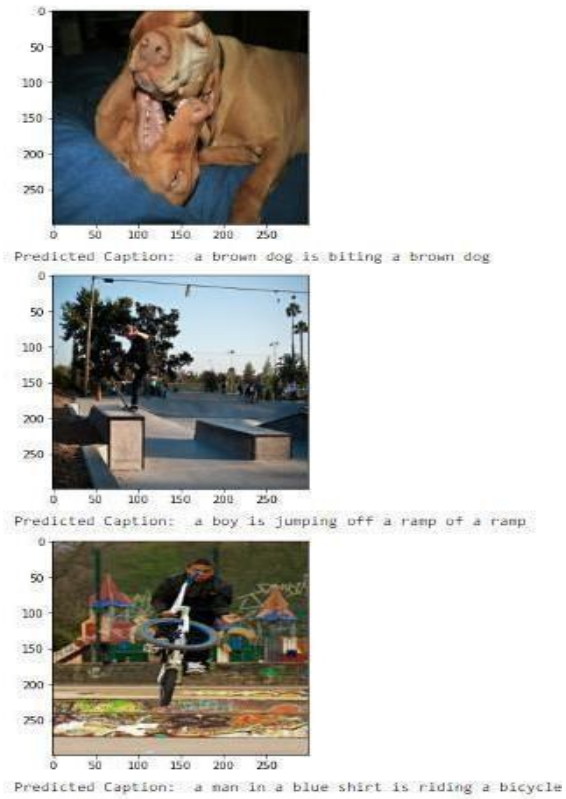


Fig.11 Efficient algorithm

Observation Of Results:

The examples above show the outcomes of picture captioning using VGG16 and InceptionV3. The Flickr8K Dataset, which included 1000 test photographs, was used to generate the results. With both CNN variations, the model was assessed using the CIDEr score. The resulting VGG16 and InceptionV3 CIDEr scores are 0.3692 and 0.3572, respectively. The training with those CNNs revealed that the InceptionV3 offers outcomes that are relatively similar to those of VGG16, but with a larger number of epochs. InceptionV3 requires 12 epochs while the VGG16 model only needed 7 epochs to produce equivalent results. After 7 epochs, the application of LSTM produced a CIDEr score of 0.39.

Sample Screenshots:



6. Conclusion

In this article, we've discussed about deep learning-based image captioning methods and provided some context for the project's conclusion. We looked at the benefits and drawbacks of various evaluation criteria and datasets. A brief summary of the experiment's findings is also provided. We provided a brief overview of several potential research directions. Despite the impressive progress made in recent years by deep learning-based image captioning techniques, it is still not possible to create high-quality captions for almost all photos. Automatic picture captioning will remain a hot topic for research with the introduction of novel deep learning network architectures. The captions for the nearly 8000 images that make up the Flickr 8k dataset that we used are also included in the text file. Despite the fact that deep learning-based image captioning methods have advanced significantly in recent years, a dependable method that can produce captions of a high calibre for almost all photos has not yet been created. Automatic photo captioning will continue to be an attractive study topic for some time to come with the introduction of innovative deep learning network designs. The future of image captioning is very promising as more people use social media every day and the majority of them post images. They will therefore gain more from this effort.

Fututre Scope

Picture captioning has become a big problem as a result of the internet's and social media's exponential rise in picture content in recent years. This study discusses the many image retrieval studies that have been carried out in the past and emphasizes the various approaches and strategies employed in the research. Future study in this field has a great deal of potential because it is challenging to extract features from photographs and calculate how similar they are. Using characteristics such as colour, tags, picture retrieval with image captioning 54, histogram, etc., current image retrieval systems use similarity calculation. Results cannot be completely accurate because these methodologies are independent of the image's context. In order to tackle this problem in the future, a full research of picture retrieval using the image context, such as image captioning, is necessary. By enhancing this work in the future to better identify classes with lower precision, more image captioning datasets can be used to train the system. This methodology may also be used with older image retrieval methods like the histogram, shapes, etc. to determine if the results improve.

References:

[1] J.Brownlee, .https://machinelearningmastery.com/develop-adeep-learning-caption-generation-model-in- python/ (accessed Jul. 09, 2020). “How to Develop a Deep Learning Photo Caption Generator from Scratch, “Machine Learning Mastery, Jun.26,

[2] “[1707.07102] https://arxiv.org/abs/1707.07102 (accessed Jul. 20, 2020). OBJ2TEXT:

Generating Visually Descriptive Language from Object Layouts."

[3] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, ArXiv181004020 Cs Stat, Oct. 2018, Accessed: Jul. 20, 2020. [Online]. Available: ,"A Comprehensive Survey of Deep Learning for Image Captioning,"

[4] G. Nishad, Medium, Mar.12, 2019. https://blog.goodaudience.com/automatic- imagecaptioning -building-an-image-caption generator-from-scratch-4bdd8744bc38(accessed Jul. 20, 2020), "Automatic Image Captioning : Building an image-caption generator from scratch !,"

[5] Sugumaran, V. R., & Rajaram, A. Lightweight blockchain-assisted intrusion detection systemin energy efficient MANETs. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-16.

[6] Kalaivani, K., Kshirsagarr, P.R., Sirisha Devi, J., Bandela, S.R., Colak, I., Nageswara Rao, J. and Rajaram, A., 2023. Prediction of biomedical signals using deep learning techniques. Journal of Intelligent & Fuzzy Systems, (Preprint), pp.1-14.

[7] Artificial Intelligence Stack Exchange. https://ai.stackexchange.com/questions/10114/whats-the-commercialusage-of-image-captioning (accessed Jul. 20, 2020). "Deep learning - What's the commercial usage of 'image captioning'?,"

[8] K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 15–29,] D. Hutchison et al., "Every Picture Tells a Story: Generating Sentences from Images," in Computer Vision – ECCV 2010, vol. 6314,

[9] Rathish, C.R. and Rajaram, A., Hierarchical Load Balanced Multipath Routing Protocol for Wireless Sensor Networks.

[10] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing Simple Image Descriptions using Web-scale N-grams," p. 9.

[11] Models and Evaluation Metrics Extended Abstract," p. 5, M. Hodosh, P. Young, and J. Hockenmaier, "Framing Image Description as a Ranking Task Data.

[12] Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jeju Island, Korea, Jul. 2012, pp. 359– 368, Accessed: Dec. 24, 2019. [Online]. Available:https://www.aclweb.org/anthology/P12- 1038, P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi, "Collective Generation of Natural Image Descriptions".

[13] 2019 5th International Conference on Advanced Computing Communication Systems (ICACCS), Mar. 2019, pp. 107–109, doi: 10.1109/ICACCS.2019.8728516, N. K. Kumar, D. Vigneswari, A. Mohan, K. Laxman, and J. Yuvaraj, "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach,".

[14]Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Aug. 2018, pp. 1–4, doi: 10.1109/ICCUBEA.2018.8697360,] C. Amritkar and V. Jabade, “Image Caption Generation Using Deep Learning Technique,” in 2018.

[15] Rajaram, A. and Palaniswami, S., 2010, September. The trust-based MAC-layer security protocol for mobile ad hoc networks. In 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM) (pp. 1-4). IEEE.

[16] 2017 International C onference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Mar. 2017, pp. 1–4, doi: 10.1109/ICIIECS.2017.8276124, P. Shah, V. Bakrola, and S. Pati, “Image captioning using deep neural architectures”.

[17] 25th IEEE International Conference on Image Processing (ICIP), Oct. 2018, pp. 1278– 1282, doi: 29 10.1109/ICIP.2018.8451558,] F. Fang, H. Wang, and P. Tang, “Image Captioning with Word Level Attention,” in 2018

[18] 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Aug. 2016, pp. 246– 247, doi: 10.1109/URAI.2016.7625747, D.-J. Kim, D. Yoo, B. Sim, and I. S. Kweon, “Sentence learning on deep convolutional networks for image Caption Generation,” in 2016.

[19] Rathish, C.R. and Rajaram, A., 2018. Sweeping inclusive connectivity based routing in wireless sensor networks. ARPN Journal of Engineering and Applied Sciences, 3(5), pp.1752-1760.

[20] Computer Science and Information Technologies (CSIT), Sep. 2017, pp. 162–167, doi: 10.1109/CSITechnol.2017.8312163, A. Poghosyan and H. Sarukhanyan, “Short-term memory with read-only unit in neural image caption generator,” in 2017.

[21] ArXiv170802043 Cs, Aug. 2017, Accessed: Jul. 20, 2020. [Online]. Available: , M. Tanti, A. Gatt, and K. P. Camilleri, “What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?,”.

[22] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning,” ArXiv161201887 Cs, Jun. 2017, Accessed: Jul. 20, 2020. [Online]. Available: http://arxiv.org/abs/1612.01887.

[23] M. Nguyen, “Illustrated Guide to LSTM’s and GRU’s: A step by step explanation,” Medium, Jul. 10, 2019. https:// towardsdatascience.com/ illustrated-guide-to-lstms-and-gru-s-a-step-bystep explanation-44e9eb85bf21 (accessed Jan. 01, 2020).

[24]“Flickr8K.”https://kaggle.com/shadab hussain/flickr8k (accessed Nov. 25, 2019).

[25] K. Papineni, S. Roukos, T. Ward,

and W. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation,” 2002, pp. 311–318.

[26] J. Hui, “Real-time Object Detection with YOLO, YOLOv2 and now YOLOv3,” Medium, Aug. 27, 2019. https://medium.com/@jonathan_hui/real-time-object-detection-with-yoloyolov2-28b1b93e2088 (accessed Jul. 20, 2020).

[27] “Yolo Framework | Object Detection Using Yolo,” Analytics Vidhya, Dec. 06,2018.https://www.analyticsvidhya.com/blog/2018/12/ practical-guide-object-detection-yolo Frame wor-python/ (accessed Jul. 20, 2020).