International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.8, Issue.6, June 2020 E-ISSN: 2347-2693

A Deep Learning Model for Image Caption Generation

P. Aishwarya Naidu[1]*, Satvik Vats[2], Gehna Anand[3], Nalina V.[4]

[1,2,3,4]Dept. of Information Science and Engineering, B.M.S College of Engineering, Bangalore, India

**Corresponding Author: 1bm16is062@bmsce.ac.in, Tel.: +91 9573560063*

DOI: https://doi.org/10.26438/ijcse/v8i6.1017 | Available online at: www.ijcseonline.org

Received: 06/June/2020, Accepted: 20/June/2020, Published: 30/June/2020

***Abstract*—** Computer vision has been an area of interest for engineers and scientists who have been spearheading in the field of artificial intelligence from the late 1960s as it was very essential to give machines or robots the power of visualizing objects and activities around them like the human visual system. The ability to visualize 2-Dimensional images and extracting features from them can be utilised for developing various applications. The involvement of deep learning has been successful in bolstering the field of computer vision even further. The abundance of images in today's digital world and the amount of information contained in them have made them a very valuable and research worthy data item. A deep learning-based image caption generator model can incorporate the areas of natural language processing and computer vision with deep learning to give a solution in which the machine can extract features from an image and then describe those features in a natural language. Thus, explaining the contents of the image in a human-readable format. This model has various applications ranging from social causes like being an aid to visually impaired to enhancing search experience of users over the web. This paper analyses the various state-of-the-art work in the field of image processing, computer vision and deep learning and presents a deep learning model that generates captions describing the images given as input to the system.

***Keywords*—** Image caption, Recurrent Neural Networks, Feature Extraction, Image Description

I. INTRODUCTION

Computer vision, Image processing and machine learning have become very crucial needs and also an economical alternative in various fields and applications. These include applications ranging from signature recognition for authorization to iris and face recognition in forensics. Also, their combination is being widely used in military applications across the world. Each of these applications has its special basic requirements, which may be unique from the others. Any stakeholder of such systems or models is concerned and wants their system to be faster, more accurate than other counterparts as well as cheaper and equipped with more extensive computational powers. All such traits from the systems are desirable as most of these systems are being used for mission-critical purposes and scope of any mistake should be very less. Such systems are required to handle the complexity of problems of the modern world like intelligent crimes, a smart city needs like smart traffic control systems, disaster control and management systems etc., thus a computer vision-based model that is unbiased and free of any prejudice towards anything or anyone is required to generate a caption describing the images given to it as input. So that such description can be used to automate existing systems like traffic control systems, flood control systems or surveillance systems, this will reduce chances of errors in such critical works and also the surveillance can be conducted 24X7 without human interaction. This paper puts forward a task of extracting features from a digital image and then describing those extracted features in a natural language. Through the localization and description of salient regions of images using LSTM a meaningful sentence in natural language will be formed that will describe images i.e. given a set of images and prior knowledge about the content find the correct semantic label for the entire image(s).

II. RELATED WORK

Kelvin Xu et al. have proposed a model for automatic caption generation which is developed by the combination of all recent work done in the field of machine translation with those done in the area of object detection using computer vision [1]. They have explored various features of an image based on attention model i.e. a model in which each word of the image description is generated focusing different areas of the image and progressively picking words from the vocabulary that best describes the area focused (attention area). The model described in this paper is also self-capable of learning to amend its gaze on important entities that are present in the image while generating the sentences of description in which every word is related and makes complete sense. They further describe two types of attention mechanism i.e. Hard Version and Soft Version. Authors have tested the accuracy of the developed model using benchmark datasets like BLEU and METEOR.

Tanti et al. have discussed a model for image caption generation that relies on neural models, however, instead of retrieving image descriptions (either partial or wholesale) they have generated new captions by using a recurrent neural network which is usually a long short-term memory (LSTM) [2]. Normally, such models use image features separated from a pre-trained convolutional neural system (CNN). For example, the InceptionV3 CNN to predisposition the RNN towards examining terms from the vocabulary so that a grouping of such terms produces captions that apply to the picture. The state-of-the-artwork puts forward two views of how the RNNs will be used, Non-memory based RNNs and Memory based RNNs respectively.

Bernardi et al. have ordered the current methodologies of caption generation dependent on how they conceptualize this issue, namely, models that give depiction a role as either generation problem or as a retrieval issue over a visual or multimodal illustrative space [3]. It gives a point by point audit of existing models, featuring their advantages and disadvantages. In addition to that, it gives an overview of the benchmark image datasets and the assessment measures that have been created to evaluate the nature of machine-generated image descriptions. It likewise extrapolates future bearings in the region of automatic image description generation. The authors finish up from the review that in contrast with the conventional keyword-based image annotation (utilizing object acknowledgement, attribute discovery, scene marking, and so forth.), automatic image description frameworks produce increasingly human-like clarifications of visual content, giving a progressively complete picture of the scene.

Vinyals et al. have developed a model called NIC in which they showcase an end-to-end neural network model that can automatically see a photo and produce a reasonable description in English [4]. NIC depends on a CNN that encodes a picture into a smaller portrayal, trailed by a recurrent neural system that creates a corresponding sentence. The introduced model is prepared to amplify the probability of the sentence given the picture. Analyses on a few datasets show the strength of NIC as far as subjective outcomes (the produced sentences are truly sensible) and quantitative evaluations, using either BLEU or ranking metrics, a metric utilized in machine interpretation to assess the nature of generated sentences. It demonstrated that, as the size of the accessible datasets for image description increments, so will the exhibition of approaches like NIC.

Kuznetsova et al. have presented an all-encompassing data-driven way to deal with image description generation by exploiting the huge measure of (boisterous) associated natural language descriptions accessible on the web and parallel image data [5]. The model retrieves a current human-made expression used to depict outwardly comparable pictures, given a query image, at that point specifically join those expressions to produce a novel description for the inquiry picture. It gives the generation procedure a role as constraint enhancement issues, altogether joining various interconnected parts of language composition for content arranging, surface acknowledgement and discourse structure. Assessment by human annotators shows that their last system produces more semantically right and phonetically engaging descriptions than two nontrivial baselines.

Siming Li et al. have presented an essential yet compelling method to manage automatically forming image descriptions given PC vision-based wellsprings of information and using web-scale n-grams [6]. Unlike most past work that layouts or recoups past content relevant to a picture, this method structures sentences totally without any planning. Test outcomes show that it is functional to deliver fundamental straightforward depictions that apply to the specific content of a picture while permitting inventiveness in the description – making for more human-like explanations than previous strategies. This strategy involves two phases i.e. (n-gram) phrase fusion and (n-gram) phrase selection.

Ryan Kiros et al. have proposed two multimodal neural language models i.e. models of trademark language that can be adjusted on various modalities [7]. An image text multimodal neural language model can be used to recuperate pictures given complex sentence inquiries, recoup phrase descriptions given picture queries and similarly produce content adapted on pictures. In contrast to a critical number of the present procedures, this technique can create sentence descriptions for pictures without the usage of configurations, organized prediction, and syntactic trees. Or maybe, it relies upon word portrayals picked up from countless words and moulding the model on high-level image features picked up from deep neural systems. They introduced two systems subjects to the log-bilinear model of Mnih and Hinton (2007) i.e. the factored 3-way log-bilinear model and the modality-biased log-bilinear model. Word portrayals and image features are discovered together by commonly setting up our language models with a convolutional network.

Yang et al. have suggested a term creation technique that explains images by anticipating the most possible verbs, propositions, nouns and verbs that make up the central sentence structure [8]. The input is the original noisy measurement of scenes and objects found in the picture using trained detectors which are state of the art. As the direct estimation of behaviour from still photographs is inaccurate, a language model is used to train the English Gigaword corpus to achieve their estimates; along with the probability of co-located nouns, scenes and prepositions. These projections are used as variables for the HMM that represent the phrase generation process, with secret nodes as phrase components and image detection as pollutants. Description of an image is the result of an incredibly complex process that involves:

1. Interpretation in Visual Space
2. Foundation in World Experience in Language Space
3. Speech / text creation.

Experimental findings indicate that this technique of integrating vision and language creates coherent and concise sentences relative to simplistic approaches that use vision alone.

Zaremba et al. have provided a brief regularization technique using Long Short-Term Memory (LSTM) from Recurrent Neural Networks (RNNs) [9]. Unluckily, for RNNs the most powerful regularization method for feed-forward neural networks does not work. As a consequence, realistic implementations of RNNs sometimes use versions that are too tiny whereas big RNNs appear to be over-fitting. Present regularization approaches have fairly minor changes for RNNs Graves (2013). Through this work, the authors show that dropout, when it is used correctly, the overfitting in LSTMs significantly reduces and assesses it on three different problems.

Barnard et al. have provided a different method for simulating multimodal data sets, based on the particular case of segmented photographs with the corresponding text [10]. There are many programs to learn about the joint distribution of visual features and words. They find in-depth the identification of terms identified with entire images (self-annotation) and referring to specific picture regions (region naming). Auto-annotation can organize and view a huge collection of photographs. Region labelling is a model of object recognition as a process of interpreting regions of images into words, which can be translated from one language to another. Learning the association between image regions and semantic correlations (words) is an excellent case of multimodal data mining, especially as it is usually difficult to apply data mining methods to picture collections.

Yao et al. have presented an Image to Text (I2T) framework that converts an image and video content to text descriptions based on image (or frame) comprehension [11]. The suggested framework is accompanied by three steps. The input images with the use of an image parsing engine are fragmented in their visual features in the first stage. In the second stage, the effects of the first stage are transformed into a textual description in the context of the Network Ontology Language (OWL). Finally, in the third stage, the OWL representation of the preceding step is transformed into semantically meaningful, human-readable and searchable text reports by a text generation engine. The I2T framework is particularly different in that it generates semantically meaningful annotations. Since the content of the image and video is converted into both OWL and text format, this framework can be merged with a full-text search engine to provide error-free content-based recovery. Users can also query images and video clips based on keywords and semblance.

Kumar et al. have used deep learning to propose an Image Caption Generator [12]. It aims at producing captions based on processes that require both image processing and computer vision for an input image. The method identifies the connections in the image between various people, objects and animals whilst capturing and turning semantic meaning into a human natural language. Regional Object Detector (RODe) is used to identify, recognize and produce captions. The method proposed is centred on deep learning to further improve the existing system. This method applies to a dataset of 8k Flickr. It created captions with a more concise meaning and detailed importance than the current generators of image captions.

Shabir et al. have deduced that, since the research community has been very interested in finding new ways to automate content-based image retrieval, they have presented an overview of some technical aspects and techniques for caption-generation of images [13]. The paper discusses briefly the description of certain new research and also the relevant points of ongoing work. There are various picture description frameworks but findings indicate that a better architecture with improved results is still needed. Four main fields are identified, where potential efforts to enhance results should be carried out.

Li et al. have incorporated the visual storytelling issue in the literature. It aims to produce coherent and concise lengthy-video stories [14]. Video storytelling poses new challenges because of the variety of the plot, the length as well as the quality of the images. The writers suggest innovative approaches for overcoming the problems. To begin with, the authors are proposing a context-aware paradigm for multimodal embedding research, they are designing a Residual Bidirectional RNN to use past and future contextual knowledge. Multimodal embedding is then used to access video clip phrases. Secondly, they suggest a Narrator model for choosing excerpts that explain the specific plot. The Narrator is intended as a reinforcement learning agent who is trained by explicitly improving the generated story's textual metric.

Xiangyang et al. have presented a system for image captioning based on scene graphs has been suggested by the writers in this paper. A few techniques have recently captured semantic ideas from photographs and converted them into high-level depictions afterwards [15]. While considerable improvement has been made, the majority of prior approaches individually handle individuals in pictures, thereby missing organized information that offers meaningful signals for image captioning. Scene graphs provide a significant volume of ordered details since they not only represent objects in pictures but often include pair-by-pair connections. CNN features from the bounding box offsets of entity instances are derived for visual representation, as well as semantic interaction features from triplets (e.g. man-eating fruit) for semantic representation, to include both image features and

conceptual knowledge in structured scenario graphs. After acquiring these characteristics, the authors present at each step a hierarchical attention-based module for learning discriminative highlights for word generation.

III. METHODOLOGY

III.I System Design

A typical recurrent neural network encoder-decoder design is used to solve the question of creating image captions. It has two aspects involved in it:

1. Encoder: A network architecture that uses an internal representation to interpret the picture data and translate the information into a fixed-length vector.
2. Decoder: A network model that reads the encoded picture and produces the output as a textual interpretation.

The merged approach integrates both the encoded image input form and the encoded text summary form that has been produced so far. A very basic decoder model then utilizes a combination of those two encoded inputs to produce the next term in the chain. The method only uses the RNN to encode the text created up to now. This distinguishes the issue of processing the input of pictures, the input of text and the mixture and perception of the encoded data.

The deep learning model development can be represented in three phases as shown in Figure 1 :

1. Photo Feature Extractor: This is an InceptionV3 layer that is pre-trained on the ImageNet dataset. The images will be pre-processed using the InceptionV3 model without the output layer and will use the derived features projected by this model as data.
2. Sequence Processor: It is a word embedding layer for managing text data, preceded by a Long Short-Term Memory (LSTM) recurrent neural network layer [16].
3. Decoder: Both the sequence processor and feature extractor produce a fixed-length vector. They are bundled together and analyzed by a Dense Layer to reach a final prediction.

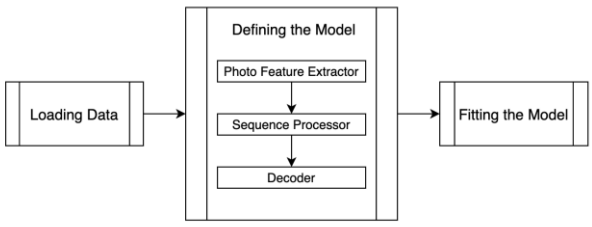

Figure 1. Development of Deep Learning Model

The data flow diagram (Figure. 2) depicts the flow of data across our proposed model which starts with preparing the photo data collected from different sources and preprocessing them to be in the expected input format. After this, the text descriptions (captions) associated with each of the images are prepared and both the text and image data are combined to develop the deep learning model using a combination of photo feature extractor, sequence processor and decoder, which is followed by training the developed model using progressive loading to ensure accuracy. Evaluation or test of the developed model is performed using the test data split from the total training set. Once the satisfactory or desired accuracy is reached the model can be used to generate captions for the new input images.

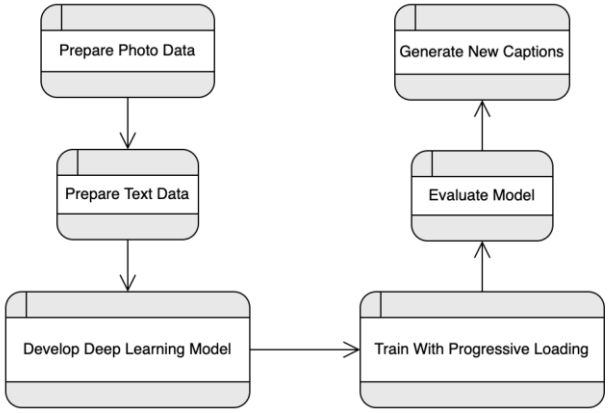

Figure 2. Data Flow Diagram

III.II Implementation

Generating a textual depiction for a given photo is a challenging artificial intelligence issue. It requires two strategies; one is to understand the content of the picture and another is a language model (from the field of NLP) which transforms the content of the picture into words organized appropriately.

III.III Data Collection and pre-processing

With the end goal of this task, MS-COCO dataset is utilized. The dataset contains 82000 pictures and each image is associated with at least 5 captions. To quicken the training speed, a subset of 30000 captions and the associated images are randomly selected to train the model. It is sensible and moderately little so it very well may be utilized to run on workstations utilizing a CPU as opposed to depending on a GPU of a powerful machine. Thus, it is conceivable to run a system that is certainly not a very good quality PC/Laptop. The pictures are split as follows:

1. Training Set (24000 pictures)
2. Test Set (6000 pictures)

Each picture is combined with five distinct captions which give away from the notable entities and events. These pictures were taken from six events, and they do not contain any notable individuals or areas, however, were physically chosen to portray an assortment of scenes and circumstances.

a little girl in a pink dress going into a wooden cabin .
a little girl climbing the stairs to her playhouse .
a little girl climbing into a wooden playhouse .
a girl going into a wooden building .
a child in a pink dress is climbing up a set of stairs in an entry way .

two dogs on pavement moving toward each other .
two dogs of different breeds looking at each other on the road .
a black dog and a white dog with brown spots are staring at each other in the street .
a black dog and a tri-colored dog playing with each other on the road .
a black dog and a spotted dog are fighting

young girl with pigtails painting outside in the grass .
there is a girl with pigtails sitting in front of a rainbow painting .
a small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
a little girl is sitting in front of a large painted rainbow .
a little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .

man laying on bench holding leash of dog sitting on ground
a shirtless man lies on a park bench with his dog .
a man sleeping on a bench outside with a white and black dog sitting next to him .
a man lays on the bench to which a white dog is also tied .
a man lays on a bench while his dog sits by him .

the man with pierced ears is wearing glasses and an orange hat .
a man with glasses is wearing a beer can crocheted hat .
a man with gauges and glasses is wearing a blitz hat .
a man wears an orange hat and glasses .
a man in an orange hat starring at something .

Figure 3. Five Captions for Each Image

A pre-trained model is utilized to decipher the contents of the photographs. The InceptionV3 model which Keras provides. The issue is, it is an enormous model and running every photograph through the system each time another language model configuration is to be tried is repetitive. Rather, by utilizing the pre-trained model the features are pre-figured and saved to a document. These features are then loaded into the model to translate a given photograph in the dataset. It is indistinguishable from running the photograph through the full InceptionV3 model (it simply happens once in advance). This improvement will consume less memory and will make preparing the models quicker. Using the InceptionV3 class, the InceptionV3 model is stacked in Keras. The last layer from the model is expelled, as this is the model which is used to anticipate a classification for a photograph. Just the internal representation of the photograph is of significance. These are the features that the model has received from the photograph. Keras also provides the devices to reshape the stacked photograph into a desired size for the model.

The MSCOCO dataset contains various portrayals for every photo and the content of the depictions requires some cleaning. Every photograph contains a novel identifier. This identifier is utilized on the photograph filename and in the content document of depictions. The depiction of the text needs cleaning. The text is cleaned in the following ways to decrease the length of the jargon of words the model needs to work with: converting all the words to lowercase, removing all punctuation, removing all words that are one character or less in length (e.g. ‘a’) and removing all words which contain numbers in them. Before the description text gets introduced to the model as an input or contrasted with the model's forecasts, it should be encoded to numbers. The initial phase is to map the words to unique integers. The Tokenizer class in Keras library provides a way to load these mapping as input, from the loaded description data.

III.IV Developing the Model

The model produces a caption given a photograph, and the caption will be produced word by word. The grouping of recently produced words will be given as input. Along these lines, there is a requirement for a 'first word' to start the generation procedure and a 'last word' to show the finish of the inscription. The string ‘<start>’ and ‘<end>’ are utilized for this. These strings are added to the stacked depictions while they are stacked. It is critical to do this so that the tokens are encoded before the text is encoded.

Before the description text gets introduced to the model as an input or contrasted with the model's forecasts, it should be encoded to numbers. The initial phase is to map the words to unique integers. The Tokenizer class in Keras library provides a way to load these mapping as input, from the loaded description data.

The text is currently to be encoded. Every depiction is split into words where the model is given a single word and the photograph which will produce the next word. Further, the first 2 words of the caption are given as input along with the picture to produce the next word. This is how the model will be prepared. For instance, the sequence "little girl running in field" is partitioned into 6 input-output sets to prepare the model as shown in Figure 4.

X1,	X2 (text sequence),	y (word)
photo	startseq,	little
photo	startseq, little,	girl
photo	startseq, little, girl,	running
photo	startseq, little, girl, running,	in
photo	startseq, little, girl, running, in,	field
photo	startseq, little, girl, running, in, field,	endseq

Figure 4. Data points corresponding to one image and its caption

Later the produced words will be linked and recursively given as input to produce a caption for a picture when the model is utilized again to produce depictions. There are 2 input arrays to the model: one for the encoded content and one for photograph features. One output from the model is the next word in the sequence which is encoded. The input text will be fed to a word embedding layer which is encoded as integers. The photo features are loaded directly to another part of the model. The model yields a prediction, which is a probability distribution over all the words in the vocabulary. The information yielded will subsequently be a one-hot encoded variant of each word, representing an idealized probability distribution with 0 values at all word positions aside from the actual word position, which has a value of 1.

The model is characterized as dependent on the "merge-model". A standard encoder-decoder RNN architecture is utilized to address the image generation issue. This included two components: Encoder and Decoder.

The merge model joins the encoded type of the text depiction produced up until now with the picture input which is also encoded. A basic decoder model is used to load these 2 encoded inputs, to produce the following word in the sequence. The methodology utilizes the RNN just to encode the text produced up until this point. This isolates

the worry of demonstrating the picture input, the text input and the combining and translation of the encoded inputs.

The Photo Feature Extractor model utilizes photograph features as input and it expects it to be a vector of 4096 components. They are then fed to the dense layer to deliver a 256-component representation of the photograph. The Sequence processor model requires the input sequences to be of predefined length (34 words). This is taken by the Embedding layer which utilizes a mask to ignore padded values. This is trailed by a LSTM layer with 256 memory units. The LSTM layer is only a particular RNN to process the sequence input i.e. partial captions. Both the input models produce a 256-component vector and use regularization as half drop out. As the model learns quickly, this is done to diminish overfitting the preparation dataset. Further, an addition operation is performed on both the input models by the decoder model which merges them into a single vector. This is then sent to a dense 256 neuron layer and afterwards to a final output Dense layer that makes a softmax prediction over the whole output vocabulary for the following word in the sequence.

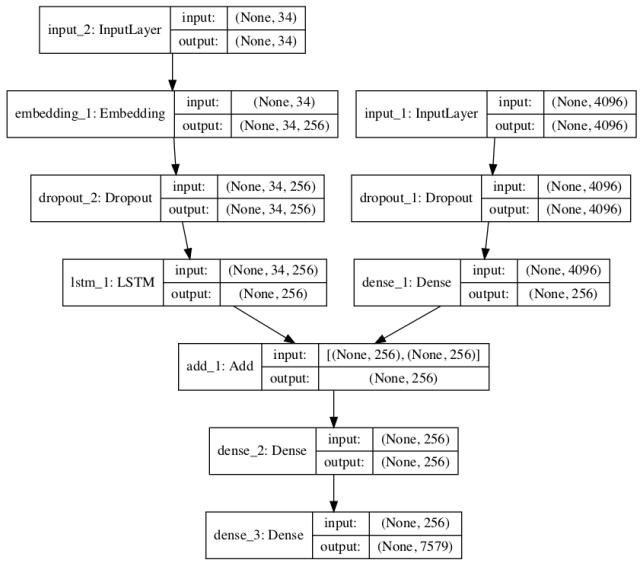

Figure 5. Model Plot

III.V Training the model

The model catches on quickly and rapidly overfits the preparation dataset. Hence, the prepared model ought to be observed on the holdout development dataset. At the point when the skill of the model improves on the development dataset towards the end of an epoch, the entire model will be saved to a document. Towards the finish of the run, the final model will be one of the saved models which give the best accuracy on the preparation dataset. This is finished by characterizing a Model Checkpoint in Keras and determining it to screen the minimum loss on the validation dataset and save the model to a record that has both validation and preparation loss in the filename. The model is fit for 20 epochs.

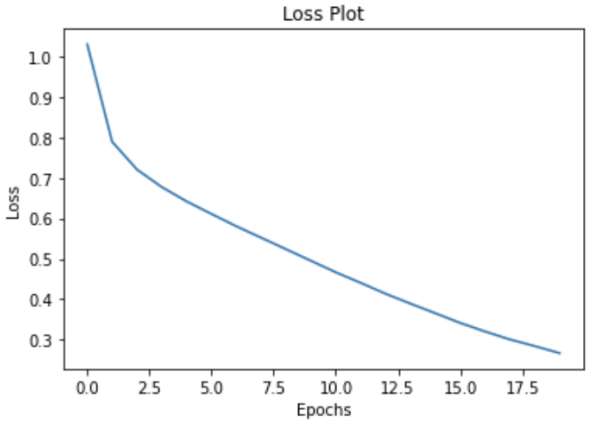

Figure 6. Loss Plot

III.VI Evaluation of the Model

Once the model is fit, it can be assessed using the holdout test dataset. The dataset contains completely 30000 photos for our use. The division of images from our dataset for different purposes while developing the model are as follows:

1. Training - 24000 data items (i.e. images & their descriptions/captions)
2. Testing - 6000 data items (i.e. images & their descriptions/captions)

The developed model is evaluated by testing the machine-generated captions for the photos in our test dataset against a standard cost function. Tests are needed to decide whether or not the model is overfitting.

Since testing the developed model involves judging the accuracy of image captions generated by the machines that are in natural language, a score or benchmark was needed to determine the accuracy of the captions. As only then the developed model can be tested whether it can generate the captions based on the training and secondly, whether those captions accurately describe images and meaningful sentences are generated by identifying various aspects of images. BLEU score was used as a benchmark to test the generated captions for images given as input from the test set [17].

IV. RESULTS AND DISCUSSION

BLEU score defines the exactness of the model. Bilingual evaluation understudy (BLEU) is an algorithm which evaluates the quality of the text that a machine has interpreted. It was one of the first to reach a high correlation with human judgment. BLEU score is always defined as 0 to 1, where 0 means that the generated image description or caption is not able to describe the image features correctly at all.

1. A numerical closeness metric of the translation which is then generated for captions generated for each of the images and measured against
2. A corpus of translations by human reference.

To measure the BLEU score, the captions are generated first for all the test images, and then the captions produced by these machines are used as nominee sentences. The candidate sentences are contrasted with the human-given captions and the candidate's average BLEU score relating to each of the references. So we measure 6000 BLEU scores for 6000 test photos using the Natural Language Toolkit (NLTK) which is a python package. We compare each description generated against all photo reference descriptions. We then calculate cumulative n-grams of BLEU scores for 1, 2 and 3.

An average of the BLEU score over 6000 test images are taken. The model's net BLEU score after 20 epoch training was calculated to be 0.37 for uni-grams, 0.18 for bi-gram and 0.11 for tri-gram which show that the model generated captions are close to human-generated captions.

From the BLEU scores, it can be seen that uni-grams have the highest score followed by bi-grams and lastly tri-grams. This is expected since, in unigrams score, the individual words are compared while in bi-grams and tri-grams, groups of 2 or 3 words are compared. The exact ordering and context might be different from the human-generated sentence which leads to its lower score compared to uni-grams. The state of the art BLEU outcomes can be obtained by raising the number of epochs and increasing the number of training data points but that will entail higher computation. A few example results are given.

Table 1. BLEU Scores

N-grams of Generated Caption	BLEU Score
1	0.37
2	0.18
3	0.11

Figure 7. A nice vase sitting near a tree holds flowers

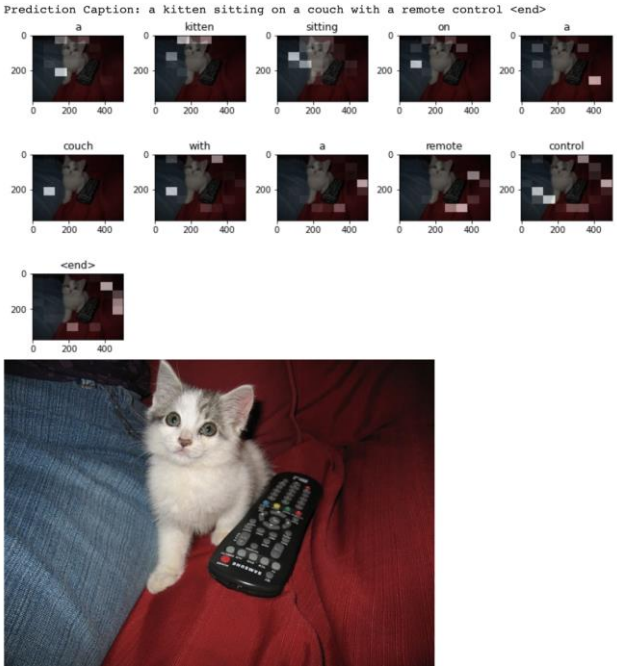

Figure 8. A kitten sitting on a couch with a remote control

V. CONCLUSION AND FUTURE SCOPE

V.I Conclusions

The deep learning model presented in this paper uses Long-Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs) and successfully demonstrates the development of an image caption generator tool, as evident from Figures 7 and 8. The obtained results for new input images show that the trained model could detect relationships between various objects in images as well as the actions/position of those objects (like surfing, standing, sitting near, etc.) (Figure 7 & 8). The BLEU score is used in the presented model to evaluate it and a suitable accuracy value was achieved for the model, the obtained scores were 0.37, 0.18 and 0.11 for 1-gram, 2-gram and 3-grams of generated captions respectively (Table 1).

V.II Future Enhancements

The fact to understand is that, although images have emerged to be a highly abundant data item across the globe and how it is highly important to analyse the image data, it is not the end. Videos are also available in huge amounts and they also carry the same amount of information and sometimes even more critical and useful information as images. Hence, there is a need for a similar deep learning model that can describe the contents of a video in a natural language.

Developing such a model would help in domains of security and military applications, real-time crowd management applications, driverless vehicle technologies and help visually impaired people etc. An extension of the present model can be developed to tackle this challenge of describing video contents in a natural language by a machine.

REFERENCES

[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio. "Show, attend and tell: Neural image caption generation with visual attention" In International conference on machine learning, pp. 2048-2057, 2015.

[2] M. Tanti, A. Gatt and K.P. Camilleri. "*What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?*" arXiv preprint arXiv:1708.02043, 2017.

[3] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, B. Plank. "*Automatic description generation from images: A survey of models, datasets, and evaluation measures*", Journal of Artificial Intelligence Research, Vol. 55, pp. 409-442, 2016.

[4] O. Vinyals, A. Toshev, S. Bengio, D. Erhan. "Show and tell: A neural image caption generator", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164, 2015.

[5] P. Kuznetsova, V. Ordonez, A.C. Berg, T.L. Berg, Y. Choi. "Collective generation of natural image descriptions", In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 359-368, 2012.

[6] S. Li, G. Kulkarni, T. Berg, A. Berg, Y. Choi. "Composing simple image descriptions using web-scale n-grams", IN Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, pp. 220-228, 2011.

[7] R. Kiros, R. Salakhutdinov, R. Zemel. "Multimodal neural language model", In International conference on machine learning, p. 595-603, 2014.

[8] Y. Yang, C. Teo, H. Daumé III, Y. Aloimonos. "Corpus-guided sentence generation of natural images", In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 444-454, 2011.

[9] W. Zaremba, I. Sutskever, O. Vinyals. "*Recurrent neural network regularization*", arXiv preprint arXiv:1409.2329, 2014.

[10] K. Barnard, P. Duygulu, D. Forsyth, .N de Freitas, D. Blei, M. Jordan. "*Matching words and pictures.*" Journal of Machine Learning Research, Vol. 3(Feb), pp. 1107-1135, 2003.

[11] B. Yao, X. Yang, L. Lin, M. Lee, S. Zhu. "I2T: Image parsing to Text Description", In Proceedings of the IEEE, pp. 1485-1508, 2010.

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

[13] S. Shabir, S. Arafat. "An image conveys a message: A brief survey on image description generation", In 2018 1st International Conference on Power, Energy and Smart Grid (ICPESG), pp. 1-6, 2018.

[14] J. Li, Y. Wong, Q. Zhao, M. Kankanhalli. "*Video Storytelling: Textual Summaries for Events*", IEEE Transactions on Multimedia, Vol. 22, Issue. 2, pp. 554-565, 2019.

[15] X. Li, S. Jiang. "*Know more say less: Image captioning based on scene graphs*", IEEE Transactions on Multimedia, Vol 21, Issue 8, pp. 2117-2130, 2019.

[16] S. Kavitha, A. Senthil Kumar, "*Long Short-Term Memory Recurrent Neural Network Architectures*", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 3, pp. 390-394, May-June 2019.

[17] Anitha Nithya R, Saran A , Vinoth R, "*Adaptive Resource Allocation and Provisioning in MultiService Cloud Environments* ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 382-387, March-April 2019.

Authors Profile

Ms. P Aishwarya Naidu is currently in her final year pursuing Bachelor of Engineering in Information Science and Engineering in B.M.S College of Engineering, Bangalore affiliated to Viveswaraya Technological University.

Mr. Satvik Vats is currently in his final year pursuing Bachelor of Engineering in Information Science and Engineering in B.M.S College of Engineering, Bangalore affiliated to Viveswaraya Technological University.

Ms. Gehna Anand is currently in her final year pursuing Bachelor of Engineering in Information Science and Engineering in B.M.S College of Engineering, Bangalore affiliated to Viveswaraya Technological University.

Mrs. Nalina V is working as Assistant Professor in the Department of Information Science and Engineering at BMS College of Engineering, Bangalore, India. She has received BE (VTU Belgaum) in Computer Science and Engineering in 2008 and M.Tech in Computer Network Engineering from Dayananda Sagar College of Engineering (VTU), Bangalore in 2010.

