Image & Video Caption Generator

[1] Prof. Namrata Khade
[1]Guide, Department of Computer Science & Engineering,
Priyadarshini College of Engineering,
Nagpur, India

[2] Princy Meshram
[2]UG Student, Department of Computer Science &
Engineering, Priyadarshini College of Engineering,
Nagpur, India

[3] Sakshi Pote
[1]UG Student, Department of Computer Science &
Engineering, Priyadarshini College of Engineering,
Nagpur, India

[4] Vaishnavi Bodhale
[2]UG Student, Department of Computer Science &
Engineering, Priyadarshini College of Engineering,
Nagpur, India

[5] Kirti Kawale
[5]UG Student, Department of Computer Science &
Engineering, Priyadarshini College of Engineering,
Nagpur, India

[6] Rutika Wadaskar
[5]UG Student, Department of Computer Science &
Engineering, Priyadarshini College of Engineering,
Nagpur, India

***Abstract* - Images are the visual memories that have a significant influence on the human encephalon, enabling us to recall fine details like a place or a specific person or object that we instantaneously record. However, some of the images are not recognised, and one needs a precise description of them in order to get a precise notion of what the image actually comprises. The process of understanding an image's context and adding appropriate captions to it makes use of deep learning and computer vision. With the use of datasets made available during model training, it comprises tagging a picture with English keywords. To train the CNN model Xception, the imagenet dataset is used. Image feature extraction is handled by Xception. The LSTM model will be modified by these extracted features to provide the picture caption. Applications that automatically try to provide captions or explanations regarding picture and video frames have a lot of promise when using deep learning-based methodologies. In the field of imaging science, captioning for both images and videos is seen as a clever problem. The application fields include general-purpose robot vision systems, automatic creation of metadata for pictures and videos (indexing), automatic development of captions (or descriptions) for images and videos for persons with various degrees of visual impairment, and many more. Numerous more task-concrete applications can benefit greatly from each of these application categories.**

Keywords - Image captioning, video captioning, Machine Learning, LSTM, neural network, image processing.

I. INTRODUCTION

Image processing has had a major impact on both science and industry and will continued to do so. Applications for it have been found in several fields, including scene comprehension and visual perception, to name a couple. Prior to the development of deep learning, the majority of researchers relied on imaging techniques that performed well on rigid objects in controlled settings using specialised gear. Deep learning-based convolutional neural networks have lately had a positive and significant impact on the area of photo captioning, allowing for considerably more versatility. In this post, we choose to highlight recent breakthroughs in the field of the picture and video labeling within the context of deep learning. Deep learning-based convolutional neural networks have lately had a positive and significant impact on the field of photo captioning, providing considerably more freedom. In this essay, we choose to highlight recent advancements in the field of picture and video annotation within the context of deep learning.

II. LITERATURE SURVEY

A. Verma, H. Saxena, et al [1], Humans have the propensity to infer significance from all they observe, alive or not. The entire situation inspired us to take this step and investigate computer vision and how it may be used with neural networks that recur to provide captions for any image. Given the recent rise in applications based on natural language processing, numerous other scholars have also worked on this idea and obtained fantastic results. It is difficult to describe a picture accurately; language structure and semantics play a significant role in grammatical structure. In order to generate effective and relevant captions by properly training the dataset, this study handles the task of caption production with an LSTM based Recurrent neural network and builds approach that relies on the same. The Flicker8k dataset was successfully used to train our model. The model's accuracy is assessed using common assessment metrics.

V. Agrawal, S. Dhekane et al [2], the procedure for creating captions for photographs is used to create sentences that describe the situation that was photographed. It locates the key aspects of the image, recognises parts of the image, and carries out a few actions. After the system has identified this data, it should next produce the most pertinent and succinct description of the image that is also semantically and syntactically sound. With the advancement of machine-learning techniques, algorithms are now able to produce text in the form of naturally occurring sentences that can accurately describe images. It is difficult for a machine to mimic human abilities to comprehend content of the image and produce descriptive text. The uses for automatic image caption creation are numerous and important.The challenge entails creating succinct captions utilising a variety of approaches, including Deep Learning (DL), Computer Vision (CV), and Natural Language Processing (NLP). This study presents a system that generates the captions using an encoder, a decoder, and an attention method. It first extracts the image's features using a pre-trained CNN called Inception V3 and then utilises a RNN entitled GRU to provide pertinent caption. The suggested model employs a Wellpositioned attention mechanism to produce captions. The model is trained using the MS-COCO dataset. The outcomes confirm that the model can reasonably comprehend photos and produce text.

C. Amritkar and V. Jabade et al [3], in intelligent machines, computer vision and natural language processing are used to automatically create an image's contents NLP. The regenerative neuronal model is developed. It is dependent on machine translation and computer vision. Using this technique, natural phrases are produced that finally explain the image. CNN and recurrent neural networks are also components of this approach RNN. The CNN model is used to extract features from images, and the RNN is used to generate sentences. The model has been taught to produce titles that, when given an input image, almost exactly describe the image. On various datasets, the model's precision and the fluency or command of the language it learns from visual representations are examined.

E. Mulyanto et al [4], computer vision research faces a hurdle with image captioning. In the Indonesian context, this work advances research on the automatic production of image captions. For unidentified photographs, a description in Indonesian phrases is generated. FEEH-ID, the first Indonesian sequence-to-sequence dataset, is the dataset that was used. Due to the lack of an Indonesian corpus for image captioning, this research is essential. Using the CNN and LSTM models, this study will compare the experimental findings in the FEEH-ID dataset with datasets in English, Chinese, and Japanese. With scores of 50.0 for BLEU-1 and 23.9 for BLEU-3, which are above average for Bleu assessment results in other linguistic datasets, the performance of the proposed model in the test set shows promise. the model for blending CNN and LSTM.

L. Abisha Anto Ignatious et al [5], The semantic tags included in the image are used to label the items that have been identified. By include these contextual labels in the captions, it improves how effectively captions describe the items. The captions are generated by the Sequence - to - sequence language model one word at a time. The faces dataset, which contains the facial images of 232 celebrities, is used by the face recognition algorithm to identify and recognise the faces of celebrities in photos. Personalized captions were created by replacing the mentions of the persons in the sentence with their names. To determine the accuracy of the generated captions, METEOR and the Bilingual Evaluation Understudy levels were established.

M. P. R, M. Anu et al [6], The optimal method for this project is the merging of CNN and LSTM the primary goal of the suggested research is to find the ideal narrative for an image. The description will be translated into the text after being obtained, and the text will then be given voice. For persons who are blind and cannot understand visuals, image descriptions are the greatest option. If their vision cannot be corrected, the descriptive can be generated as a speech output using a voice-based image caption generator. Image processing will become a prominent research area in the present, mostly used to save a person's life.

S. Li and L. Huang et al [7], captioning images is crucial but challenging. The current picture caption mostly uses an encoding and decoding structure, with CNN serving as the primary image feature extractor in the encoder and LSTM serving as the primary decoder. In the current encoding and decoding structure, the attention mechanism is also frequently exploited. However, the convolutional neural networks and recurrent neural networks-based image caption models now in use are not very accurate in extracting relevant information from images and have issues like gradient explosion. This research suggests a context-based image caption-generating approach to solving these issues. The technique uses SCST and LSTM for captioning, followed by SCST and context coding for feature extraction. The efficiency of the suggested technique is shown by the experimental findings.

T-Y LIN et al [8], the dataset has been statistically analysed in-depth by the authors and compared to PASCAL, SUNI, and ImageNet. Then, utilising a Deformable Parts Model, we present baseline functional testing for segmentation detection and bounding box results. The collection includes images of 95 different objects kinds that a 4-year-old could recognise with ease. Our dataset was produced using unique software platforms for category recognition, instance spotting, and instance segmentation, with a total approximately two million tagged instances in 33800 photos.

A Karpathy et al [9], the authors describe a model that can produce summaries of images and their regions in natural language. Our method makes use of databases of sentence-described images to discover the cross-modal correspondences between language and visual data. Our alignment methodology is built on a cutting-edge combination of bidirectional machine learning algorithms over phrases, convolutional neural networks over image areas, and a structured objective that aligns the two modalities via multimodal embedding.

P J TANG et al [10], the new layers are used to refine and reserve the LSTM model. A weighted average method is utilised to combine the final predicted probability for all of the Softmax functions that are supplied with the respective classification layers during the test. Experimental results on the Flickr30K, MSCOCO and datasets show that our model is efficient and performs better on a number of assessment metrics than other techniques of the same kind.

III. METHODOLOGY

Here in this project, we suggest a probabilistic and neural framework for extracting descriptions from photos. Recent developments in lemmatization have demonstrated that, with a strong sequence model, it is possible to directly maximise the probability of the right translation given an input language in an "end-to-end" manner, both for training and inference. These models employ a recurrent neural network, which converts the variable-length input into a fixed-dimensional vector and then "decodes" it to produce the required output sentence. Therefore, it makes sense to employ the same strategy, using the same concept of "translating" an image into its description instead of a source language input sentence.

Therefore, we suggest using the following formulation to directly maximise the likelihood of the proper description given the image:

$$\theta \star = \arg\max \theta \; X \; (I,S) \log p(S|I; \theta)$$

S its correct transcription. Since S represents any sentence, its length is unbounded. Thus, it is common to apply the chain rule to model the joint probability over S0 to SN, where N is the length of this particular example as

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

When, out of convenience, we discarded our reliance on using stochastic gradient descent, we optimise the average of the log probabilities over the whole training set for the training example pair (S, I) at training time.

A. Convolution Neural Network Model

CNN is a type of deep learning model that can classify and identify images and objects by processing data as a 2D matrix. By reading or scanning the image from top to bottom or left to right, the details of the image can be retrieved. The information in the image can be obtained by analysing the features. You can also use transformed images [11][12][13].

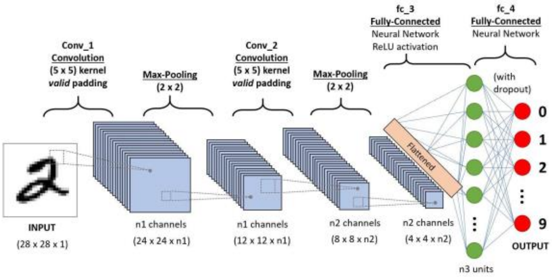


Fig. 1 CNN Model Architecture for image Classification

We employ a convolutional neural network to represent pictures (CNN). They are currently cutting-edge for object identification and detection since they have been extensively used and investigated for picture tasks. Our particular choice of CNN produces the greatest result at the moment in the ILSVRC 2014 classification competition [14] and makes use of a novel batch normalization method. By using transfer learning, they have also been demonstrated to generalize to other tasks including scene classification [15]. A word embedding model is used to represent the words.

B. Long Short-Term Memory (LSTM) Model

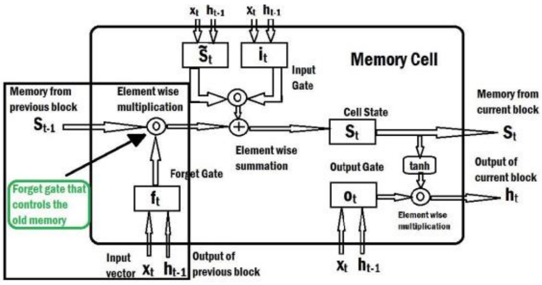


Fig. 2 LSTM Memory Lock Architecture

A cell called cell C, which is managed by three gates, is part of the memory block. The recurrent connections are represented by the blue arrows: the predicted word at time t 1 is fed back into the Softmax for word prediction in addition to the memory output m at time t via the three gates and the cell value via the forget gate.

After viewing the image and the words that come before it as defined by p(St|I, S0, St1), the LSTM model is trained to predict each word of the sentence. It is helpful to think of the LSTM in its unrolled form, where a duplicate of its memory is made for each word and image in a sentence such that all LSTMs use the same set of parameters, and where the output from the LSTM at time t is sent to the LSTM at time t. (shown in Fig. 2). In the unrolled version, all recurrent connections are converted to feed-forward connections. The unrolling technique is as follows, if we indicate by I the input image and by S = (S0,...,, SN) a true sentence characterizing this image:

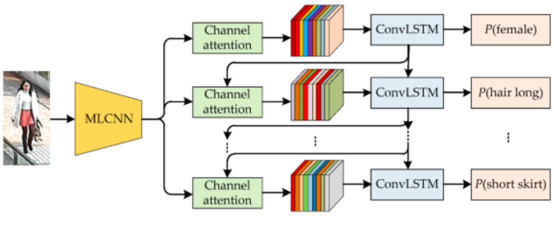


Fig. 3 Combination of LSTM Model & CNN Model

C. Working

Python was used to create this piece. Numerous Python libraries were utilised for the implementation, including Keras, which included the VCG net responsible for item detection, and TensorFlow, created by Google, which is used to build DNN by running a number of basic algorithms.

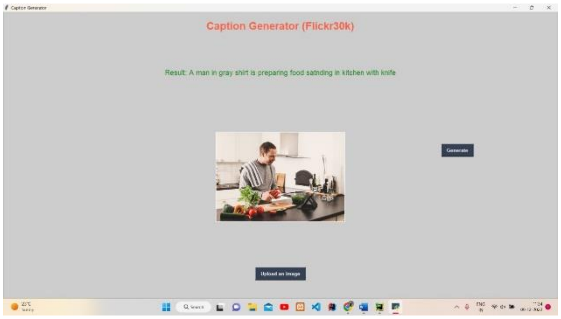
1. The initial stage was to begin by extracting features using CNN and applying different layers with our model upon this training set for which we had a corresponding text that mapped to the activity occurring in the picture. We were able to successfully extract the characteristics of each image by employing multiple layers, including conv2d, max pooling, dropout, and activation functions.

2. We used Google's word2vec model to extract features from additional testing photos as that was required. Word2vec information It is a Google model that enables us to convert our words into numerals for use in processing that involves words. It is more frequently utilised in research relating to NLP, CNN, and LSTM.

3. When a word is entered into word2vec, it is transformed into a vector with set dimensions. The context of the words in the sentence is now maintained using the LSTM layer because the meaning of a sentence might change depending on the position of each word inside it.

The dataset that we have used for our research work is available online named as “flicker8k”. The dataset was preprocessed and made fit for further evaluation and working. It comprised of 800 images, out of which we used for training and for testing in a ratio of 60:30. At the time of feature extraction, we had a total of 500 parameters out of which 475 were trained successfully and 25were non-trainable parameters. The general confusion matrix was used for analyzing the system performance. This matrix contains result of all models with their predictions. A total of 150 iterations were executed over a standard batch size of 6.5

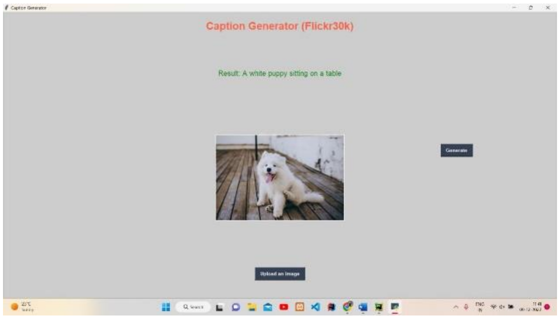
IV. RESULT



Caption Generator (Flickr30k)

Result: A man in gray shirt is preparing food satnding in kitchen with knife

Fig. 4.1 Result 1



Caption Generator (Flickr30k)

Result: A white puppy sitting on a table

Fig. 4.2 Result 2

V. CONCLUSION

By creating a model based on LSTM-based CNN adept at screening and obtaining information from any provided image and converting it to a single-line phrase based on a natural language of English, we have overcome past limitations that were experienced in the field of image captioning. Although it is acknowledged that avoiding the overfitting of data can be challenging, we are happy to have succeeded in doing so. The algorithmic core of various attention methods received the majority of attention. Hereby, we may claim that we were successful in creating a model that is a vastly superior version of every other image caption generator that was previously available.

VI. FUTURE SCOPE

As depicted in the image, we are able to implant a camera in the shoe's front face to capture real-time environment video and obtain a means to connect it wirelessly to the blind person's Bluetooth in-ear. The only difference now that this Arduino equipment is being used is that the annotations will be generated in a dynamic environment and made to be played in the blind person's Bluetooth device so that he can cross with more caution. This will undoubtedly reduce accidents and mishaps specifically involving blind people.

IV. REFERENCES

[1] A. Verma, H. Saxena, M. Jaiswal and P. Tanwar, "Intelligence Embedded Image Caption Generator using LSTM based RNN Model," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 963-967, doi: 10.1109/ICCES51350.2021.9489253.

[2] V. Agrawal, S. Dhekane, N. Tuniya and V. Vyas, "Image Caption Generator Using Attention Mechanism," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579967.

[3] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697360.

[4] E. Mulyanto, E. I. Setiawan, E. M. Yuniarno and M. H. Purnomo, "Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset," 2019 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Tianjin, China, 2019, pp. 1-5, doi: 10.1109/CIVEMSA45640.2019.9071632.

[5] L. Abisha Anto Ignatious., S. Jeevitha., M. Madhurambigai. and M. Hemalatha., "A Semantic Driven CNN – LSTM Architecture

for Personalised Image Caption Generation," 2019 11th International Conference on Advanced Computing (ICoAC), Chennai, India, 2019, pp. 356-362, doi: 10.1109/ICoAC48765.2019.246867.

[6] M. P. R, M. Anu and D. S, "Building A Voice Based Image Caption Generator with Deep Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 943-948, doi: 10.1109/ICICCS51141.2021.9432091.

[7] S. Li and L. Huang, "Context-based Image Caption using Deep Learning," 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2021, pp. 820-823, doi: 10.1109/ICSP51882.2021.9408871.

[8] T-Y LIN, M MAIRE, S BELONGIE et al., "Microsoft COCO:Common Objects in Context", Proceedings of the 2014 Euro-pean Conference on Computer Vision, pp. 740-755, 2014.

[9] A KARPATHY and F-F. LI, "Deep visual-semantic alignments for gen-erating image descriptions", Proceedings of the 2015 International Conference on Computer Vision and Pattern Recognition, pp. 3128-3137, 2015. [Google Scholar]

[10] P J TANG, H L WANG and K S XU, "Multi-objective layer-wise optimization and multi-level probability fusion for image description generation using LSTM", Acta Automatica Sinica, vol. 44, no. 7, pp. 1237-1249, 2018. [Google Scholar]

[11] Suma, V. "A Novel Information retrieval system for distributed cloud using Hybrid Deep Fuzzy Hashing Algorithm." JITDW 2, no. 03 (2020): 151-160.

[12] Manoharan, Samuel. "Supervised Learning for Microclimatic parameter Estimation in a Greenhouse environment for productive Agronomics." Journal of Artificial Intelligence 2, no. 03 (2020): 170-176.

[13] Tanwar Poonam & Rai Priyanka," A proposed system for opinion mining using machine learning, NLP and classifiers", IAES International Journal of Artificial Intelligence (IJ-AI) Vol. 9, No. 4, December 2020, pp. 726~733 ISSN: 2252-8938, DOI: 10.11591/ijai.v9.i4.pp726- 733.

[14] H. R. Tavakoli, R. Shetty, B. Ali, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in Proceedings of the IEEE Conference on International Conference on Computer Vision, pp. 2506– 2515, Venice, Italy, October 2017.

[15] I.Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In NIPS, 2014.