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 visual memories that have a profound impact on the human encephalon, allowing us to recall specific information about a location, a certain person, or an object that we have captured in an instant. However, not all of the images are recognised, and a precise description of them is required to comprehend exactly what the image consists of. Deep learning and computer vision are used to understand the context of an image and classify it appropriately. It comprises labelling a photograph with English keywords by utilising datasets made available during model training. The CNN model Xception is trained using the imagenet dataset. Xception is in charge of image feature extraction. These recovered features will be used to modify the LSTM model in order to generate the image caption. Applications that automatically try to provide captions or explanations relating to photo and video frames have a lot of potential when using deep learning-based approaches. Captioning for photographs and videos is regarded as a key issue in imaging science. Among the application domains are general-purpose robot vision systems, automatically creating captions (or explanations) for images and videos for people with varying degrees of visual impairment, and many other application sectors. Each of these application groups would benefit greatly from additional task-concrete applications.

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

I. INTRODUCTION

Both research and business have been and will continue to benefit greatly from image processing. It has uses in many different fields, including scene comprehension and visual perception, to name a couple. Prior to the invention of deep learning, the majority of researchers depended on imaging methods that performed best when used in controlled environments with specialised equipment on rigid objects. Deep learning-based convolutional neural networks have lately had a positive and considerable impact on the field of photo captioning, allowing for significantly greater variation. We will cover current developments in the field of deep learning-based image and video classification in this post.

It is quite challenging for people to describe a situation in a picture or video clip. In order to create machines with this capability, computer scientists have been exploring ways to integrate the science of understanding human language with the technology of automatically extracting and interpreting visual input. Image captioning and video captioning need more work than image recognition due to the added challenge of identifying the objects and events in the image and creating a succinct meaningful statement based on the information discovered.

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

i. IMAGE CAPTIONING

The automatic description of an image's content is a fundamental challenge in artificial intelligence that integrates computer vision with natural language processing. Earlier techniques generated annotations first. There are numerous methods for captioning images. Prior to deep neural networks (DNNs), models were retrieved-based [20] or template-based [29]. Deep neural networks are being used in new approaches. There are two stages to creating an automatic caption for an image.

First, the information from the image must be retrieved and stored in a feature vector. Using deep learning models, this level focuses on visual recognition. The feature vector is then passed on to the second stage. The second stage is caption generation, which is the process of describing what was retrieved in a grammatically correct natural language sentence.

There are three primary parts to this approach. Input/output word embeddings make up the first and last components, respectively. However, their CNN-based technique uses masked convolutions while the middle component in the RNN example uses LSTM or GRU units. This part is feed-forward and has no recurrent functionality. Their CNN with attention (Attn) produced similar results. The activations of the conv-layer were also used to test an attention mechanism with attention parameters.

A. Convolution Neural Network Model

A typical Deep Learning neural network architecture in computer vision is the convolutional neural network (CNN). A computer can comprehend and analyse visual data or images thanks to the field of artificial intelligence known as computer vision.

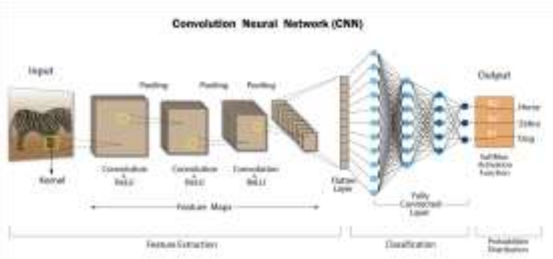


Fig. 1 CNN Model Architecture for image Classification

Convolution neural networks can learn complicated objects and patterns thanks to their input layer, output layer, numerous hidden layers, and millions of parameters. The given input is sub-sampled using convolution, pooling, and an activation function. All of these are hidden layers that are only partially connected, and at the very end is the fully connected layer that yields the output layer. The output keeps its original form and size like the input image.

B. Long Short-Term Memory (LSTM) Model

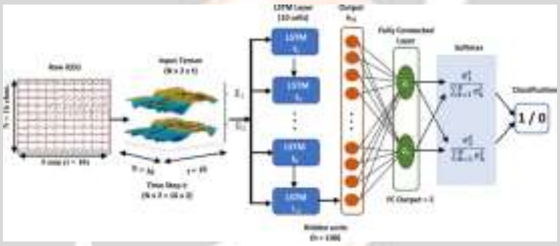
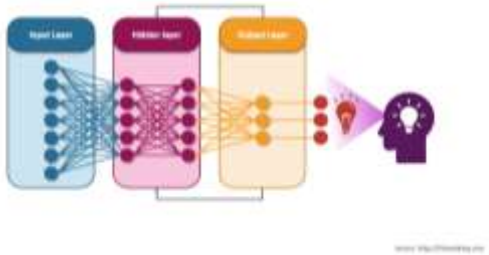


Fig. 2 LSTM Neural Network Architecture

An artificial neural network called Long Short-Term Memory (LSTM)[1] is used in deep learning and artificial intelligence. LSTM features feedback connections as opposed to typical feedforward neural networks. Such a recurrent neural network (RNN) can analyse whole data sequences, such as audio or video, in addition to single data points, like images. Because of this feature, LSTM networks are perfect for handling and forecasting data. For instance, LSTM can be used for applications like speech recognition, machine translation, speech activity detection, robot control, video games, healthcare, and unsegmented, connected handwriting recognition.

The probability of each word in the lexicon is the LSTM's output. Sentences are produced via beam search. A heuristic search strategy called beam search expands the most promising node within a small group to explore a graph. We construct sentences using both beam search and k-best search. It resembles the time synchronous Viterbi search a lot. The process preserves only the top k sentences after iteratively choosing the top k sentences out of all the candidate sentences up until time t.

C. Recurrent Neural Network



An artificial neural network that employs sequential data or time series data is known as a recurrent neural network (RNN). These deep learning algorithms are included into well-known programmes like Siri, voice search, and Google Translate. They are frequently employed for ordinal or temporal issues, such as language translation, natural language processing (nlp), speech recognition, and image captioning.

D. Implementation

In order to determine the alignments between the text segments and the associated represented areas in the image, the suggested model developed a multimodal embedding space employing the two modalities. A CNN trained on the ImageNet dataset with 200 classes was used to identify the objects, and RCNN, or regional convolutional neural networks, were utilised in the model to detect the object region. The research advises using BRNN or Bidirectional Neural Networks for language modelling since it most accurately captures the inter-modal interactions between the sentence's n-grams. The word vectors are obtained through 300d word2vec embedding.

The datasets that we used for our research are titled "flicker8k" and "flicker30k" and may be found online. The dataset was preprocessed and prepared for further analysis and study. It had 800 pictures, 60:30 of which were utilised for training and 30 for testing. We had a total of 500 parameters at the time of feature extraction, of which 475 were successfully trained and 25 were non-trainable. For analysing system performance, the general confusion matrix was used. This matrix comprises the outcomes of all models as well as their predictions. A total of 150 iterations were performed across a batch size of 6.5.

ii. VIDEO CAPTIONING

Most individuals find it easy to describe a video in normal language, while machines find it difficult. From a methodological standpoint, categorising the models or algorithms is tricky because it is difficult to establish the contributions of the visual aspects and the linguistic model used to the final description.

While it may appear that video captioning faces a problem similar to that of image captioning, another significant obstacle is the lack of datasets that include rich video descriptions. This is because videos are significantly more expensive and difficult to gather and annotate, not to mention that the task of describing visual content is becoming increasingly ambiguous. The production of video captions using learning-based techniques has only had sporadic success due to the widespread use of a number of traditional benchmark datasets in early efforts. However, they are typically brief in terms of the quantity of videos and level of narration, and straightforward in terms of both the literary and visual material. In order to properly utilise deep learning in producing human-level video description, various large-scale highly annotated video captioning datasets have been created up till recently.

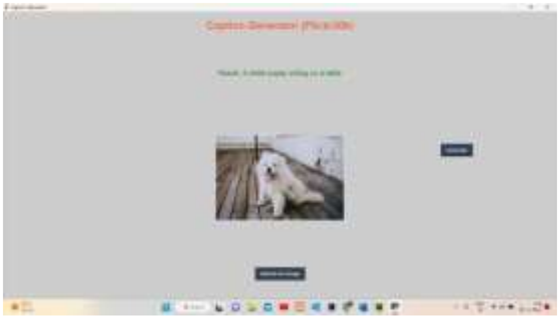
Methods for video captioning are evaluated using a variety of datasets. Here, we only name a few of them and group them into five categories based on the topic of the videos: Individuals, General Topics, Social Media, Cooking, and Movie.

One of the earlier efforts, TACoS Dataset [10], features films of several actions in the cooking area in an indoor environment. It has a total of 18,227 video-sentence pairs spread across 7,206 distinct time intervals. The TACoS-Multi Dataset is an extension to the dataset that includes a paragraph description for each temporal segment, however the limitation remains that the setting is closed-domain and too simplistic for learning.

The first open world datasets is the Microsoft Video Description Corpus (MSVD) [2], often known as the Youtube Dataset in early publications. It is a collection of YouTube videos that Mechanical Turk users selected by asking them to choose brief snippets that each represented a particular activity. Each clip lasts between 10 and 25 seconds as a result, with a fairly stable semantics and a simple temporal structure.

Similar to M-VAD, MPII Movie Description Corpus (MPII-MD) [12] is a current large-scale movie description dataset. It has over 37,000 movie clips from 55 movies with audio descriptions (ADs) and roughly 31,000 from 49 Hollywood films.

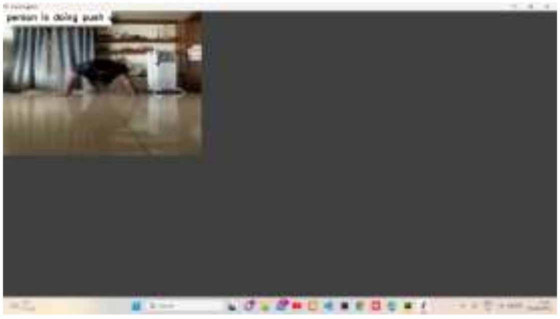
IV. RESULT



Caption Generator (Flickr30k)

Result: A white puppy sitting on a table

Fig. 4.1 Result 1



person is doing push

Fig. 4.2 Result 2

V. CONCLUSION

1.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.
2.With the help of captions, we developed a top-down saliency strategy that can be utilized to comprehend the intricate decision-making processes in image and video captioning without requiring changes like the addition of explicit captions.

layers of focus. Our method is more accurate than current approaches while yet maintaining decent captioning performance. A wide range of encoder-decoder configurations can be understood using the model's generality.

VI. FUTURE SCOPE

In order to capture real-time environment video and obtain a way to wirelessly connect it to the blind person's Bluetooth in-ear, we are able to implant a camera in the front face of the shoe, as shown in the illustration. The annotations will be generated in a dynamic environment and designed to be played in the blind person's Bluetooth device so that he can cross with greater caution now that this Arduino equipment is being used. By doing this, accidents and incidents notably involving blind persons will surely decrease.

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 Interna-tional 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. [29] J. Hockenmaier, M. Hodosh, A. Lai, and P. Young. Introducing Novel Similarity Measures for Semantic Inference over Event Descriptions: From Picture Descriptions to Visual Denotations.
14. 2014, 2:67–78, Transactions of the Association for Computational Linguistics. 5 \s[30] R. Fergus and M. D. Zeiler. Convolutional networks: Comprehension and Visualization. 2014, pages 818-833 in European Conference on Computer Vision
15. [31] X. Shen, J. Brandt, Z. Lin, J. Zhang, and S. Sclaroff. Topdown Neural Attention through Backproper Excitation. 2016, 1, 2, and 7 [32] European Conference on Computer Vision A. Oliva, L. A., B. Zhou, A. Khosla, and A. Torralba. To enable discriminative localization, learn deep features. Computer Vision and Pattern Recognition IEEE Conference.