Automatic Video Caption Generator

Mr.CH Sathyanarayana[1] , S Harish[2] , V Rohith[3] M Vishith Reddy[4]
[1]Assistant Professor, ECE, Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad
[2,3] ECE, Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad

ABSTRACT

In today's high-tech environment, where everything is based on research done recently and on which we have built, video processing has grown in relevance for a variety of reasons. Additionally, it has become crucial that other film genres, such as social and educational videos, find a place in our daily lives and surrounds. In recordings that have been tagged with video captions, you can get data by looking for specific words or phrases. Additionally, it can help the blind by explaining what is happening around them, and its threat identification and caption decoding capabilities may be useful for military operations and surveillance.

Keywords: Convolutional neural network , recurrent neural network, LSTM.

INTRODUCTION:

In several recent applications, deep learning has transformed computer vision. A machine can match or even surpass human performance in a variety of tasks, including picture classification, object identification, and video segmentation, by learning deep features and representations. But in the past twoyears, literature has paid a great deal of attention to activities like captioning images and videos, which continue to be difficult. Captioning is extremely challenging due to the nature of a video stream's high temporal dependencies, several scenes in a complicated video, and a variety of objects and events. Nevertheless, a number of systems and techniques have been put forth that significantly advanceresearch in video description**.** Building on prior achievements, this thesis work creates strong captioning frameworks that can provide captions for both straightforward and difficult videos automatically. The act of captioning videos has become more and more common in recent years. Short form video has solidified its position as an essential component of our daily lives with the rise of video sharing websites like YouTube, Twitch, and Instagram Reels. In fact, more than 500 million people use Facebook every day to watch videos, according to Forbes! 72 hours' worth of fresh video being uploaded to YouTube every minute. Artificial intelligence (AI) video solutions are now necessary due to the popularity boom of videos

LITERATURE SURVEY

Using natural language to describe a quick film is a simple task for humans but a challenging one for robots. Automatic video description is the process of using computer vision algorithms to identify the occurrences of different elements in a video. The entire scene is made up of a variety of various aspects, such as people and objects, human activities, human-human interactions, and more. Then, in order to communicate this information in a straightforward and understandable way, Natural Language Processing (NLP) techniques must be applied. In order to meet the increased demand for interpretation and description of images and videos, Computer Vision (CV) and Natural Language Processing (NLP) have teamed together recently. As well as collaborative seminars between NLP and CV conferences, there are special issues of journals devoted to language in vision. Automatic video description has a wide range of applications, including monitoring, automated video subtitling, and human-robot interaction. By automatically constructing and reading out film descriptions or by employing speech synthesis to provide vocal descriptions of the environment, it may be utilized to help persons who are visually impaired. Currently, this cannot be done at all due to the enormous cost and time involved with manual processes. Videos of sign language may also be described in this way using everyday language.

Examples of how video description might be used to generate written instructions for humans or service robots include assembling furniture, installing CDROMs, brewing coffee, or changing a flat tyre. Technology for video description has developed to the point where a variety of applications are now possible. In the near future, we anticipate being able to communicate with robots in the same manner we do with humans. If video description develops to the point of being able to analyze real-world events and translate them into spoken words, it will be considerably simpler for service robots and mobile apps to understand human behaviors and other happenings in the real world. For example, they might answer a user's question about what to make for dinner or where they left their wallet. Using these gadgets, a worker might be reminded of any jobs or procedures that are missing from a typical operation. According to Talk the Walk, a recently released dataset, natural language conversations between tourists and guides may now be utilized to help tourists uncover previously un discovered destinations on maps utilizing perception, action, and interaction modeling.

PROPOSED METHODOLOGIES

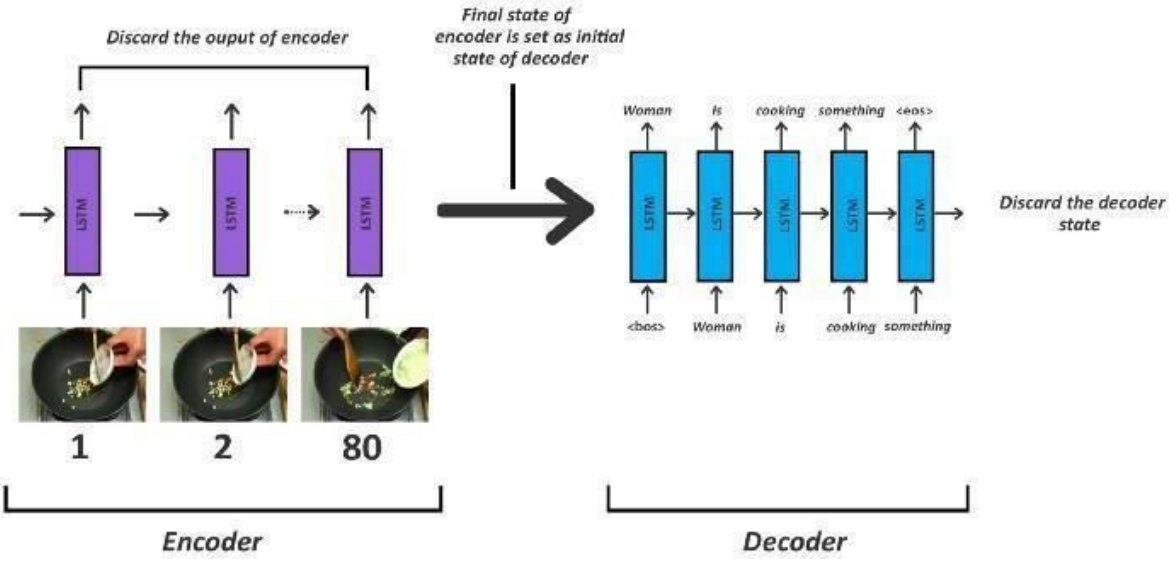
To understand this post, you must be familiar with Keras, LSTM/RNN concepts, and the fundamentals of encoder-decoder architecture. Depending on how long the movie is, different numbers of frames willbe taken. Just 80 frames from each video are utilized because of this. Each of the 80 frames that are analyzed by a pre-trained VGG16 yields 4096 attributes. Arrays of (80, 4096) features are piled on top of each other. There are 80 frames and 4096 features total, with each frame having its own set of features. Here, you can see the model VGG16 being loaded. Each frame is fed into the model, which outputs an array of numpy arrays holding information on each of the 80 frames in each movie. Since these attributes have already been collected from the data collection, we can move on to the following step. For this analysis, I used the MSVD data set from Microsoft. You may download this dataset right now. This dataset contains short YouTube videos that have been manually labelled for training and testing.

MODEL FOR TRAINING

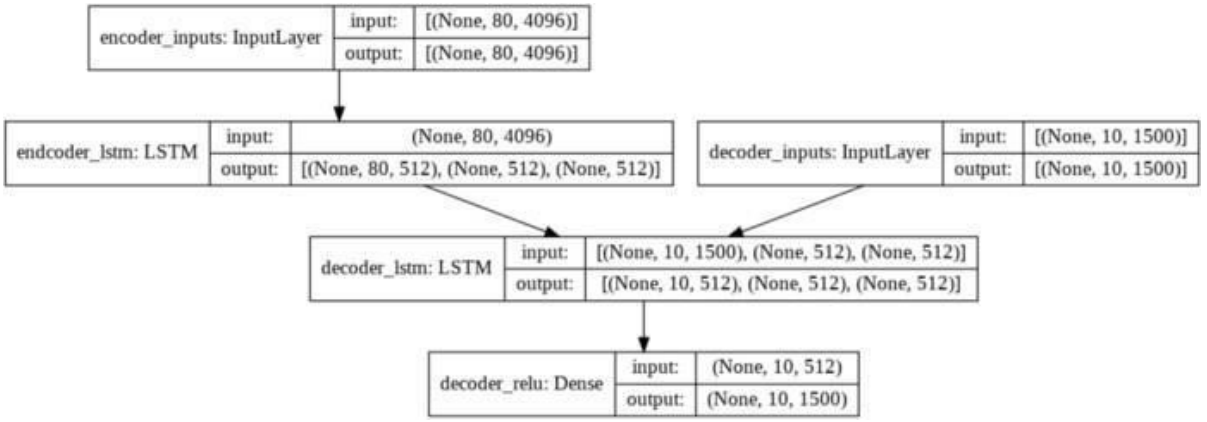
Encoder-decoder architectures are frequently employed for text production problems. We'll use this sequence-to-sequence architecture in this area as well because it's a necessary component of our approach to produce text. To find out more about this architecture, see this article. In this architecture, the decoder cell's starting state is always the encoder cell's final state. The video features will be transmitted using a video encoder, and the subtitles will be received via a decoder. Now that we are aware of what we are doing, let's examine how we will use the encoder-decoder paradigm.

A video is exactly what? We can categorize it as a succession of images, right? For jobs involving sequences, we often utilize RNNs or LSTMs. In this case, an LSTM will be employed. Visit this page to learn more about LSTMs. Now that the encoder will be an LSTM, let's have a look at the decoder. The decoder will provide captions for the video. Given that captions are merely a collection of words, we willalso use LSTM for the decoding process.

Journal of Engineering Sciences
Vol 13, Issue 06, JUNE/2022
ISSN NO:0377-9254



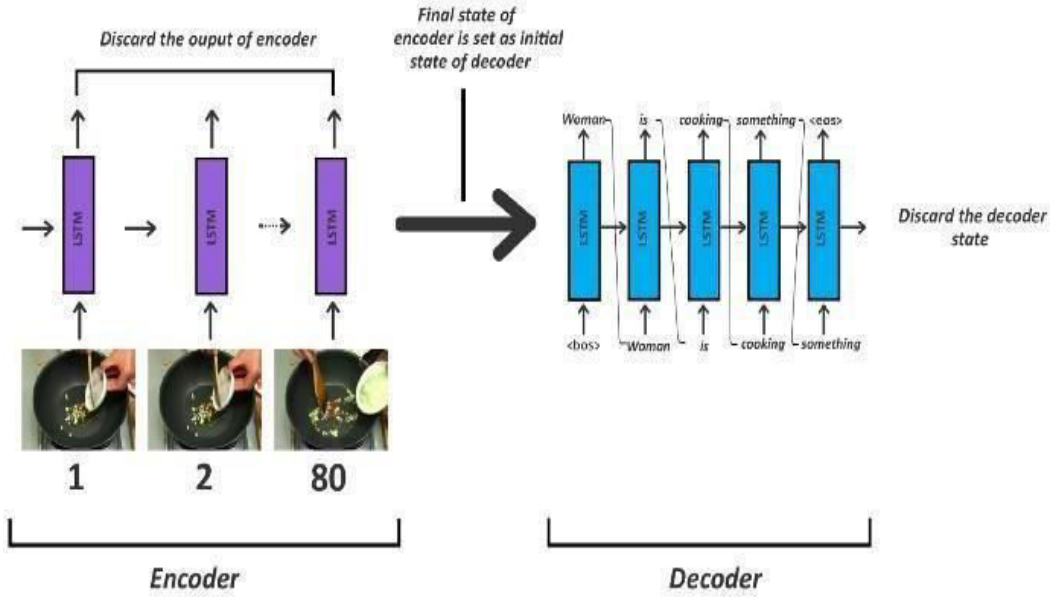
Here, the encoder's first LSTM cell is given the properties of the first frame from the image. Thefeatures of the second frame are then added, and so on until frame eighty. All other encoder outputs are ignored because in this scenario, we only care about the encoder's final state. The initial state of the decoder LSTM is now the end state of the encoder LSTM. In the first decoding stage of this decoder, thetext is entered into an LSTM bos>. One word at a time, up until eos>, is provided as training data.



Journal of Engineering Sciences
Vol 13, Issue 06, JUNE/2022
ISSN NO:0377-9254

MODEL FOR INFERENCE

The encoder-training decoder's and testing models are independent, in contrast to other neural networks. We don't keep the entire model after training. The decoder part is not preserved with the encoder model. Take a moment to examine the inference model. We will start by utilising the encoder model. Each of the 80 frames' characteristics are incorporated into the model. This model has not been altered by training. The encoder model in this situation allows us to make predictions. We won't pay attention to the encoder's other outputs because we only care about the final output state. The decoder utilises the end state of the encoder and the bos> token as its beginning state in order to predict the next word. As you can see, the model should accurately recognise the token as a woman if it has been properly trained. Remember how the captions in training were always missing the following word? the incoming data. Since there are no captions in this cell, the next word indicates the output of that cell's LSTM. The condition of the previous cell and the output lady are then transmitted into the following cell. In this way, the word that follows is foreseen. This keeps happening until the model forecasts eos accurately. We don't need any more forecasts because the sentenceis ended.



JES Journal of Engineering Sciences
Vol 13, Issue 06, JUNE/2022
ISSN NO:0377-9254

RESULTS AND DISCUSSIONS

Let me now show you some additional findings from the testing data, keep in mind that these outcomes were obtained via the use of algorithms that are notoriously greedy.

a man is performing on a stage

a man is mixing ingredients in a bowl

a cat is playing the piano

a man is spreading a tortilla

JES Journal of Engineering Sciences
Vol 13, Issue 06, JUNE/2022
ISSN NO:0377-9254

Now, it would be wrong to show only the appropriate results. Here are some of the not so correct results.

a man is riding a bicycle

The model confuses a bike to a bicycle.

a dog is making a dance

Somehow the model confuses the cat to a dog and instead of swinging the paws the model thinks of it as making a dance. This caption grammatically does not make a lot of sense.

CONCLUSION

More data shuffles are one technique to improve training. Video content from a variety of sources may be included. Automatically describing videos with natural language text enables more efficient search and retrieval, can aid visual understanding in the medical, security, and military applications, and can even be used to describe pictorial content to the visually impaired. Attention models learn to select the most relevant segments that associate with the text.

REFERENCES

[1] L. Baraldi, C. Grana, and R. Cucchiara. Hierarchical boundary-aware neural encoder for videocaptioning. arXiv preprint arXiv:1611.09312, 2016.

Journal of Engineering Sciences
Vol 13, Issue 06, JUNE/2022
ISSN NO:0377-9254

[2] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 190–200. Association for Computational Linguistics, 2011.

[3] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem. Daps: Deep action proposals for action understanding. In European Conference on Computer Vision, pages 768–784. Springer, 2016.

[4] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities fromvideo images based on concept hierarchy of actions. International Journal of Computer Vision, 50(2):171–184, 2002.

[5] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. InProceedings of the IEEE International Conference on Computer Vision, volume 1, page 6, 2017.

[6] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkila, and ¨ N. Yokoya. Learning joint representations of videos and sentences with web image search. In European Conference on Computer Vision, pages 651–667. Springer, 2016. 8

[7] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition, pages 1029– 1038, 2016.

[8] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridgevideo and language. 2016.