Real Time Video Captioning Using Deep Learning

Amrit Mohapatra (222IT004)
Information Technology
National Institute of Technology,
Karnataka, Surathkal, India 575025
amritmohapatra.222it004@nitk.edu.in

Neeraj Kumawat (222IT024)
Information Technology
National Institute of Technology,
Karnataka, Surathkal, India 575025
neerajkumawat.222it024@nitk.edu.in

Prof.Dinesh Nayak
Information Technology
National Institute of Technology,
Karnataka, Surathkal, India 575025
kiranmanjappa@nitk.edu.in

***Abstract* - In today's world of highly developed technology, where almost everything is the result of research conducted in the not-too-distant past and where we have built upon those discoveries in order to advance at an extremely rapid rate, the processing of video has emerged as a topic of utmost significance for a number of different reasons. It is also necessary to make sure that other kinds of films, such as surveillance, social, and informational videos, find their way into our day-to-day lives as well as our surroundings. Through the use of video captioning, a number of different items may be recognised, the video can be summarised and described, and data can be searched. Additionally, it may aid blind individuals by describing the events that are happening around them. Additionally, it can help in military operations and surveillance by identifying threats and assisting weapons and troops in destroying them. Finally, it can benefit persons who are visually impaired. The video encoder as well as the caption de-coder framework are both used by the video caption generator. In this study work, we have explored two models: the first model is referred to as the Hierarchical model, and the second model is referred to as the Multi stream hierarchical Boundary Model. In this approach, the hierarchical model is paired with guided captioning. In order to display a video, a hierarchical model can fundamentally gather clip-level temporal data from clips at set time steps. When defining clips in a video, the Multi-stream Hierarchical Boundary model uses a fixed hierarchy model in conjunction with a soft hierarchy model and the assistance of intrinsic feature boundary cuts. On the other hand, the Steered captioning model is the attention model, and in this model, visual parameters are used to lead an attention model to the appropriate locations in the video. A parametric Gaussian attention is also included in this study paper's discussion. The restriction of soft attention approaches is they must use video streams with a fixed length, which is eliminated by the usage of Gaussian attention techniques.**

Keywords - VGG16, Video captioning,Image Captioning

I. Introduction

The field of machine learning is quite expansive and may be categorised as a subset of artificial intelligence. Deep learning is a subtype of machine learning that can learn from data that is both unstructured and unlabeled. This ability distinguishes it from traditional machine learning. Deep learning is a relatively new development that has brought about significant changes in the field of computer vision. Machines can give performance that is comparable to or better than that of human beings in object recognition, image classification, and video segmentation by utilising the features and representations of deep learning; however, further development is still required in areas such as image and video captioning.

When there are intricate sceneries in films and several types of things present, which can occasionally pose problems for captioning, the task of creating captions for those videos becomes quite challenging. In addition, there is an issue with video captioning, and that issue is the nature of the video stream having a high degree of temporal dependencies. Due to the fact that several models and architectures have been offered for overcoming all of these challenges, research on video captioning is progressing further than ever before, which is inspiring individuals to continue research even further. This research was also motivated by a number of other research projects that have been carried out in the recent past, and we have constructed strong captioning frameworks on top of those research projects such that it is now able to generate captions for films that are both basic and complicated.

II. CONVOLUTIONAL layer

The convolutional layer, which is responsible for the majority of the network's computational processes, is the most vital component of a convolutional neural network (CNN). This layer is helpful for extracting both high-level and low-level characteristics, so it's important to keep that in mind when you use it. High-level features include input from the picture and edges, whereas low-level features include colour, grade orientation, edges, and so on. High-level characteristics also include input from the image. In order to achieve full convolution, the input is first multiplied by a large number of extremely small sliding windows, also known as kernels or filters, which are subsequently employed by each and every Convolutional Layer. To illustrate, the size of the input RGB CI-FAR-10 picture is 32 x 32 x 3, so let's take it as an example. The very beginning layer has a size of 5 x 3 x 3, and it has 16 filters and a stride 1 of 1. The output of this layer is 28 x 28 x 16, and it is to be convolved with the picture in the example. The picture has zeroes inserted around its edges so that the dimensions of its output may be modified to match those of its input. In the convolutional layer, there are three major parameters that may be adjusted to regulate the output volume. These are the number of zero padding, the depth, and the stride. The stride is what controls the sliding of the filter across the input, and the size of the stride has an inverse proportional relationship to the size of the output spatially. The depth of the analysis is determined by the number of filters

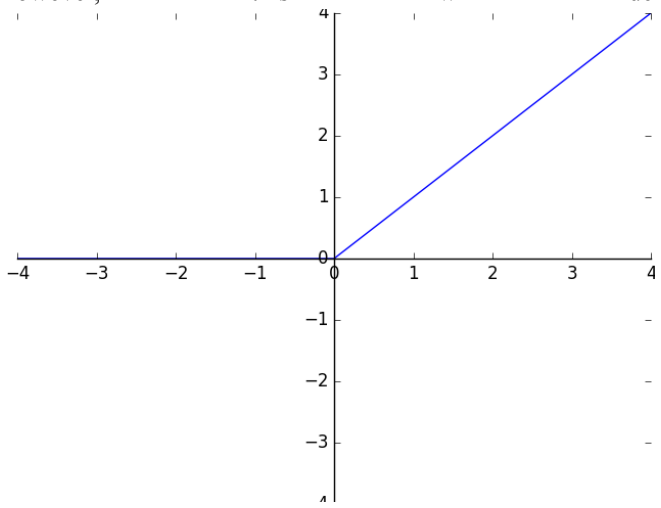
used, and each filter imparts unique characteristics upon the incoming data. In order to reduce the overall size of the input volume, zero padding is utilised.

IV. POOLING LAYER

The Pooling layer is responsible for reducing the dimensionality of the network as part of its duty. They are interspersed at regular intervals between each convolutional layer in the network. Additionally, the incorporation of crucial characteristics like rotational and positional invariant in this layer makes effective model training much easier to maintain. This layer is highly helpful in this regard. There are two distinct varieties of the Pooling Layer: the first is known as Max-Pooling, while the second is referred to as Average Pooling.

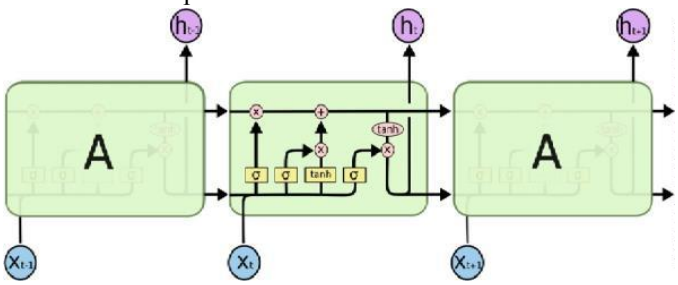
In addition to performing dimension reduction, max pooling assigns the greatest possible value to the portion of the picture that is being obscured by the kernel. Noise Suppressant is another name for this substance because of the role it plays in reducing background noise.

Now let's speak about the Average pooling algorithm. This algorithm accomplishes dimensionality reduction and also delivers the average values from the parts of the picture that are covered by the Kernel. The performance of the former type of pooling, known as max-pooling, has been seen to be significantly superior to that of the later type, known as average-pooling, after monitoring both types of pooling. The Convolutional Layer and the Pooling Layer are the two components that make up the Convolutional Neural Network's k-th layer. The number of Convolutional Layers can have a number of different counts depending on how complicated the image is. The number of convolutional layers can be raised to better capture the low-level information of the picture; however, this will dem



and a greater amount of processing resources.

With the assistance of neural networks, non-linear activation function inserts are carried out between each layer. Following the preliminary presentation of neural networks, various non-linearity functions were discovered. Sigmoid functions, also known as the hyperbolic tangent function, rectified Linear functions, and tanh functions are all examples of different types of functions. The last one is the one that is used in modern designs the most, despite the fact that it does not provide functions f



or sophisticated calculations such as sigmoid or tanh functions.

III. VIDEO CAPTIONING

the operation of Neural Networks applied to films with the purpose of directly modelling the languages. The process of video categorization is becoming commonplace in the creation of deep neural networks.

For the initial stage of the video captioning process in recurrent neural networks, the mean pool feature is used to perform the video representation.

An alternate method for it is called encoderdecoder. In this method, l frames are encoded first, one at a time, and then sent to the LSTM first layer of two layer, where the variable length is l.

The natural language phrase is decoded from the latent representation using one word at a time, and the output from each time step is fed into the LSTM second layer while the mean time passes.

This has been demonstrated by S2VT.

At first, the attention mechanism was suggested for use in the context of video captioning, and it was deployed there.

On a Text-generating recurrent network, this allows for the selection of appropriate temporal segments of video that has been conditioned.

Over the various geometric components, emphasis has been shown.

In this process, they guide the word generation to investigate certain regions of the image by making use of the output from the most recent convolution layer.

One of the benefits of choosing an image area that is proportionate to the needed text is represented by reinforcement learning, which is a method that requires strict attention.

To improve the quality of the picture captioning, semantic attention is employed. This is accomplished by picking a different list of word characteristics, which, as was previously said, helps to produce better captions through the addition of tags or video attributes.

It is challenging to acquire contents and rich attributes for videos that can also classify objects with activities since tag selection or attribute is not instructed together with the

language model. These challenges make it tough to obtain contents and rich attributes for videos.

With the use of recurrent neural networks, video captioning is now being expanded beyond the sentence and word level to construct paragraphs.

With the assistance of hierarchical recurrent networks, a video may be encoded inside of an embedding before it is used to generate words.

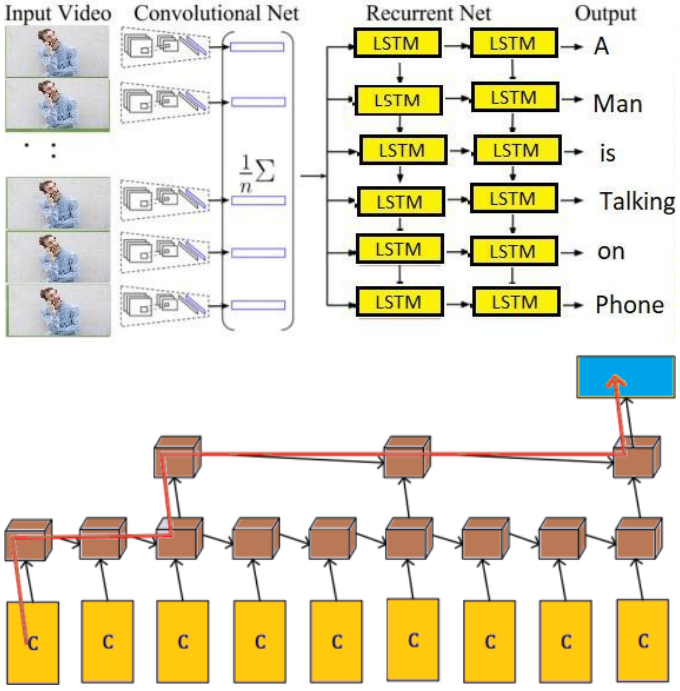
Additionally, learnable characteristics are expanded in order to apply attention across many phases, including regional, global, and local.

All of these techniques are only useful when large amounts of data that are matched with video sentences are available.

Within the context of the explanation of knowledge transfer for picture captioning derived from a set of images and independent language. This hypothesis is only tangentially motivated by the study that was conducted to enhance the producing quality of captions using visual concepts that are independent of sentences. On the other hand, the classic paradigm of soft attention that we use focuses on teaching the independent temporal video idea.

IV. Encoder Decoder models

In the starting stage of this project, videos are first represented with a mean pooled feature for the purpose of captioning of video utilising Recurrent Neural Networks (RNN), as is seen in figure 5. The LSTM layer input is the mean of all of the features in the array [5].



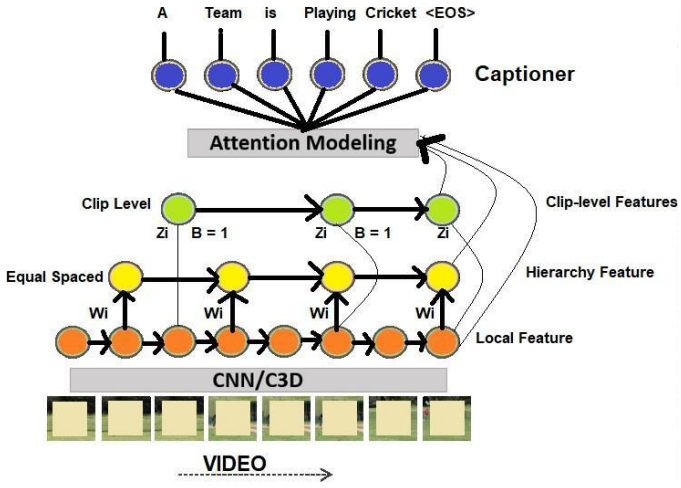
$$\Phi(V) = \frac{1}{n}\sum_{i=1}^{n} v_i [5]$$

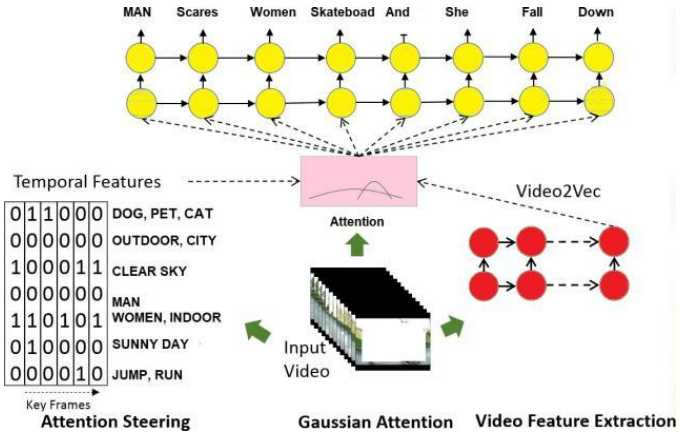
IV. Methodology

The input for the encoding stage will be the video stream, and the first layer will receive the local features in the form of x1, x2,..., xn. After that, these two layers will provide us the output in the form of a two-vector sequence. The first one has a uniform gap between it, [w1, w2, w3 wp]. and second one is clip levels [z1, z2, z3, z*q*] The output value M is given to the first vector sequence that is evenly spaced. The first layer sent these M outputs to the second layer, where M = n / k, where n is the feature count (input) and k is the intended stride value. The second clipped level vector sequence will use the information on the short boundary, which will be led by a learnt vector. This will be done by making use of the information. The cosine distance is the basis for this newly learnt vector, which is:

$$z_i = y_i.(\Delta(i,j).W_{yd} + b_{yd}) [6]$$

where Wyd represents the learned weights, b yd represents the learned bias, and yi represents the output of the first layer at each time step. In Figure 8, you can see how the video will be encoded by combining the clipped level vector characteristic with the evenly spaced characteristic. We will supply a combination of frame level, uniformly spaced sequence, and clipped level (Detected Boundaries) vector sequence as input to the caption decoder. At each time step, the model will adjust the boundary weights in order to bring out the nuances of the situation.





VI. Conclusion

In recent years, improvements in procedures for the production of video captions have been made possible by works in a variety of deep-learning techniques. These breakthroughs have proven to be a landmark in terms of the accuracy of video captioning. The textual description of a movie that is cut up into still photos can make image retrieval more efficient, which is based on the content of the images themselves. In this article, a hierarchical structure is shown for entering a video, and both the properties of the video and the duration of the movie are employed as features to grab the attention of the viewer. In addition, multi-stream captioning is being discussed in this work. This type of captioning is able to create captions for videos that are either basic or complicated. This study has other potential uses, including applications in areas like as security and surveillance, the military, medicine, and assistance for those who are visually impaired, amongst others.

References

[1] Pan, P., et al. Hierarchical recurrent neural encoder for video representation with application to captioning. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[2] Olah, C., Understanding lstm networks. GITHUB blog, posted on August, 2015. 27: p. 2015. 2016

[3] Zilly, J.G., et al., Recurrent highway networks. arXiv preprint arXiv:1607.03474, 2016

[4] Karpathy, A., et al. Large-scale video classification with convolutional neural net- works. in Proceedings of the IEEE conference on Computer Vision and Pattern Recog- nition. 2014

[5] Venugopalan, S., et al. Sequence to sequence-video to text. in Proceedings of the IEEE International Conference on Computer Vision. 2015

[6] Dong, J., et al. Early Embedding and Late Reranking for Video Captioning. in Pro- ceedings of the 2016 ACM on Multimedia Conference. 2016. ACM.