Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

Real Time Video Captioning Using Deep Learning

Prof. Sandeep Samleti
Army Institute of Technology
Pune, Maharashtra

Ashish Mishra
Army Institute of Technology
Pune, Maharashtra

Alok Jhajhria
Army Institute of Technology
Pune, Maharashtra

Shivam Kumar Rai
Army Institute of Technology
Pune, Maharashtra

Gaurav Malik
Army Institute of Technology
Pune, Maharashtra

Abstract:- In this world of advanced technology where everything is devel- oping at a very fast pace, video processing has become extremely important for various reasons. It has also become important so that various kinds of videos including surveillance, social and informa- tional videos get themselves into day-to-day life as well as into our environment. By video captioning various objects can be identified, video can be summarized and describes, data can be searched. Also, it can help blind people by describing the events happening around them as well as it can help in military operations and surveillance by detecting threats and help weapons and soldiers to destroy them. Video caption generator uses video encoder as well as caption de- coder framework. In this research paper we have discussed the two models, first one is Hierarchical model and second one is Multi stream hierarchical Boundary Model. The Hierarchical model is combined with steered captioning. Hierarchical model can basi- cally capture clip level temporal features from clips at fixed time steps to show a video. A fixed hierarchy model is taken with a soft hierarchy model with the help of intrinsic feature boundary cuts in Multi-stream Hierarchical Boundary model to define clips in a video whereas Steered captioning model is the attention model in which visual parameters are used to lead an attention model to appropriate locations in video. In this research a parametric Gauss- ian attention is also discussed. Fixed length video streams which are required by soft attention techniques is a limitation which is removed by Gaussian attention techniques.

1 INTRODUCTION

Machine Learning is a very vast subject and also subset of Artificial intelligence. Deep Learning can be described as subset of Machine learning, that is capable of learning from unstructured and unlabeled data. In recent times Deep learning has drastically changed the world of Computer Vision. By using the features of deep learning features and representations, machines can give comparable or better performance than human beings in object recognition, image classification and video segmentation, but still there are developments needed in segments like image and video captioning. Video captioning becomes very difficult as there are complex scenes in videos and diverse kinds of objects are present around which sometimes causes problem for captioning. Also, there is a problem with video captioning that is the video stream with nature of high temporal dependencies. Challenging all these difficulties various models and architectures have been proposed due to which research in video captioning is going further and motivating people to carry researches forward. This research also got motivation from various researches carried out in recent past and we have built upon them robust captioning frameworks by which generation of captions for simple as well as complex videos is made possible.

2 CONVOLUTIONAL NEURAL NETWORK

Deep-Learning algorithm include Convolutional Neural Networks that taken image as inputs and identify various objects and also tell differences between various objects present in the image. Convolutional Neural Networks (CNN) are that type of network that emphasis mainly on data that has grid-like topology. CNN are adapted in numerous architecture that has architecture that includes image recognition or we can say object detection and they are successful in such practical application projects. CNN can overcome the limitations of any other conventional methods already present in image-related tasks with the help of its small filter size and its ability to understand the in-depth representation of input. We need to study convolution, pooling and activation functions in order to get to know of CNN’s functioning in a better way.

2.1 Convolutional Layer

Convolutional layer which accounts for most computational operations is the most important part of Convolutional Neural Network (CNN). This layer is useful for extraction of both high-level features as well as low-level features. High-level features has edges and input from image where as low-level features include color, grade-orientation and edges, etc. To get full convolution, input element-wise is multiplied to numerous numbers of tiny sized sliding windows (kernels or filters) which are used by every Convolutional Layer. Let us take an example, 32 x 32 x 3 is the size of the input RGB CI-FAR-10 image. 5 x 3 x 3 is the size with 16 filters and stride 1 of very initial layer, the output of 28 x 28 x 16, in which the image in example is to be convolved. The image is padded with zeroes so that output dimension can be made similar to the input dimension. To control the output volume there are three hyper parameters in convolutional layer namely, number of zero padding, depth and stride. The sliding of filter through the input is controlled by the stride, size of stride is inversely proportional to size of output spatially. Number of filters chosen is the depth, here some new features are learnt by filter from the input. To save size of input volume, zero padding is used.

2.2 Pooling Layer

Pooling layer has the job of reduction of the dimensionality of the network. They are inserted between each convolutional layer at regular intervals. Also, this layer is very much useful in maintaining

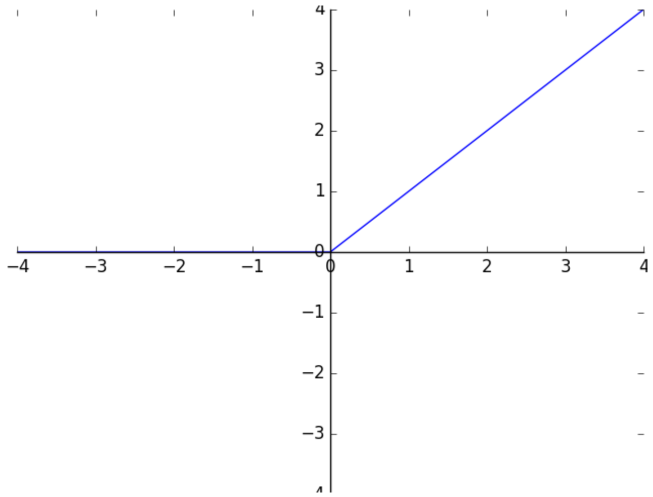

Figure 1: ReLU function

effective model training by using important features that includes rotational and positional invariant. Pooling Layer is of two types, the first one is Max-Pooling and the other one is Average Pooling. Max pooling gives highest value of that part of image which is camouflaged by kernel and also performs dimension reduction. Due to its noise suppressing property it is also called Noise Suppressant. Now talking about Average pooling, it performs dimensionality reduction and also returns that average values from the image which are covered by Kernel. After observing both types of pooling it has been observed that performance of the former i.e., max-pooling is much better than the latter i.e., average-pooling. The k-th layer of Convolutional Neural Network consists of Convolutional Layer and of Pooling Layer. The count of Convolutional Layer can be varied on the complexity of the image. To capture the low-level details of the image the number of Convolutional Layers can be increased but it requires more computational power

2.3 Activation Functions

No-linear activation functions are inserted between each layer with the help of neural networks. After the first introduction of neural networks many non-linearity functions. There are functions which are Sigmoid functions, tanh (hyperbolic tangent) function and rectified Linear functions. The last one is the most popular one in current architectures and does not include complex computations functions like sigmoid or tanh functions.

3 RNN (RECURRENT NEURAL NETWORKS)

Recurrent Neural Networks (RNN) is a form of neural network. The RNN mainly deals with the data which are in sequence, i.e

Sequential Data. RNNs process data with temporal dependencies whereas CNNs process static data at a single time along with matrix

of values like images. Recently RNNs are widely used in variety of applications like language moulding, image captioning and speech recognition. In this project we will use LSTM and RHM(Types of RNN).

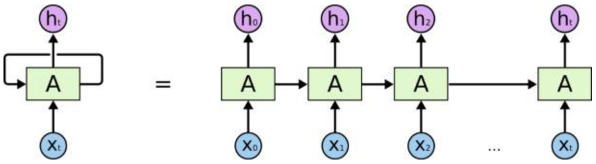

Figure 2: Expanded form of RNN [2]

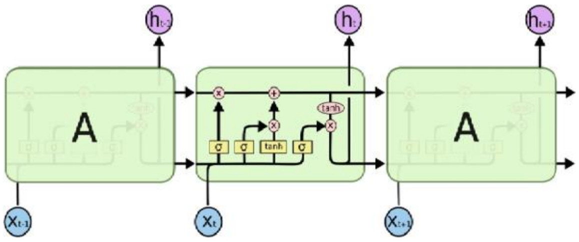

Figure 3: Inside LSTM module [2]

3.1 Vanilla Recurrent Neural Network

Vanilla Recurrent Neural Network is an example of simple RNN. RNNs attempts to learn conditions of information groupings. In Recurrent Neural Networks, next step input depends on previous computation

If at any time z input is χ_z then the invisible state h_t at given time z is given by:

$$h_z = f(U\chi_z + Wh_{z-1})$$

[10] where f - non-linear function, Then output will be calculated at step t. Now if we have to guess the upcoming word in the given sentence, we will quantify the probability of words from the whole collection of words:

$$O_t = softmax(Vs_t)$$

Now network in figure can be illustrated as:

3.2 LSTM (Long Short Term Memory) Network

Long-Transient Memory (LSTM) engineering is proposed by Hochreiter et al for tackling evaporating and detonating inclinations issue. It deals with cell states and gates. There are different types of gates which is used to monitor the stream of information with in the cell. The name of doors are neglect entryway(Forget gate), input entryway(Input gates) and yield entryway(Output gates). The forget gate tells whether the cell state information will be ignored or not. When output value of sigmoid function is 1, it means that it should be kept else if it is 0, the forget it. Where f_t is calculated by:

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)\,[10]$$

where σ - sigmoid function, U_f, W_f, b_f - learnable weight matrices and bias,

h_{t-1} - past secret territory of LSTM cell.

The input gate i_t tells whether the information is stored in the cell state. The sigmoid function that decides value to be updated is:

$$i_t = (W_i.h_{t-1} + U_i x_t + b_i)\,[10]$$

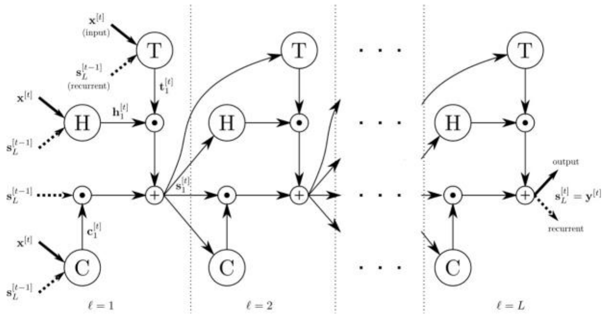

Figure 4: RHN layer inside the recurrent loop [3]

Memory cell state C_t at time any t can be calculated as:

$$C_t = \tanh(W_c.x_t + U_c h_{t-1} + b_c),\ fourth$$

The Memory cell state C_t is updated and now:

$$C_t = i_t * C_t + f_t * C_{t-1},\ fourth$$

Above equation plays major role in solving vanishing gradient problem in RNNs. Finally, output is given by the output gate:

$$O_t = (W_o x_t + U_o h_{t-1} + b_o)\ h_t$$
$$= O_t * \tanh(C_t)$$

3.3 Recurrent Highway Network

Apart from vanishing and exploding gradients to create depth for the preparing of repetitive organizations by placing different layers to create depth is also a problem with traditional RNNs. To solve this problem, Recurrent Highway Network is introduced by Zilly et al [3] which helps in train deeper model with less variables. In a RHN cell there are many expressway (Highway) layers each with two entryways(gates) the change(Transform) and convey(Carry) door. Now at any given time t, assume input is $x^{[t]}$ then output is given by:

$$s_l^{[t]} = c_l^{[t]} \odot s_{l-1}^{[t]} + t_l^{[t]} \odot h_l^{[t]}\ [10]$$
$$h_l^{[t]} = \tanh(W_H.x^{[t]}.I_{l=1} + R_{H_l}.s_{l-1}^{[t]} + b_{H_l})\ [10]$$
$$t_l^{[t]} = \tanh(W_T.x^{[t]}.I_{l=1} + R_{T_l}.s_{l-1}^{[t]} + b_{T_l})\ [10]$$
$$c_l^{[t]} = \tanh(W_C.x^{[t]}.I_{l=1} + R_{C_l}.s_{l-1}^{[t]} + b_{C_l})\ [10]$$

In this paper, We will use a type of Recurrent Highway network to check the against LSTM performance.

4 VIDEO CAPTIONING

Progess in Deep Learning video domain and its sub parts i.e Video captioning is progressed with progress in image domain part. Previously video captioning depend on withdrawing semantic contents like verb,subject,object and also linking it with optical elements. To get probability we will form a Factor Graph Model and then to the receive best mixture of verbs, object and subject to make a sentence template. Context and activity of previous work is limited to small vocabulary of activities and objects.Video-sentence pair in now available with rich language, recent research have shown the working of Neural Networks on videos to directly model the languages. Video classification is now widespread in Deep neural network construction.

In the beginning video representation is done by mean pool feature for video captioning in recurrent neural networks.Encoder-decoder is an alternative approach for it where l frames are encodes first,one at a time to the LSTM first layer of two layer, where variable length is l.Natural language sentence is decoded from that latent representation with one word at a time, one time step output is feeded to the LSTM second layer in the mean time.S2VT has been showing this.

In the beginning attention mechanism was proposed and used in the context of video captioning.On Text generating recurrent network this permit the choice of applicable temporal segments of video conditioned.Over the parts geometric attention is shown in.In which they escort the word generation to explore the particular parts of image by using last convolution layer output.A bonus of selecting the image area proportional to the required sentence is represented by reinforcement learning which is a hard-attention mechanism.To enhance the image captioning semantic attention is used by selecting separate list of word attributes have mentioned that create better captions by inserting tags or video attributes.Its difficult to get contents and rich attributes for videos that can also categorize objects with actions as tag selection or attribute is not instructed along with the language model.

Just now,at sentence and word level video captioning is widen to generate paragraph with the help of recurrent networks.Before generating words video can be encode in an embedding with the help of Hierarchical recurrent networks .Learnable parameters are also increased to apply attention over multiple stages like regional,global and local.

All these methods are only applicable when video sentence paired data is present on big scale.In the explanation of knowledge transfer for image captioning from set of image and independent language. This theory is weakly inspired from the research of improving the generating quality of captions with visual concepts of sentence independent. In opposite to this our model of traditional soft attention teaches independent temporal video concept for attention.

Our progress is more over from video captioning from model of soft attention.Using Gaussian spread over video length we can augment the attention mechanism by parameterizing.To remove dependency from model on video duration we use sigma and mean values of distribution By this multiple sentence can be generated on per video on temporal data augmentation is allowed.Gaussian filters are limited to activity classification and for word generation the use of attention is limited because of its equally spaced attention filters.

4.1 Encoder-Decoder models

In the beginning phase, videos are represented with mean pooled feature for captioning of video using Recurrent Neural Networks (RNN) as shown in figure 5.The average of all feature array is LSTM layer input[5]

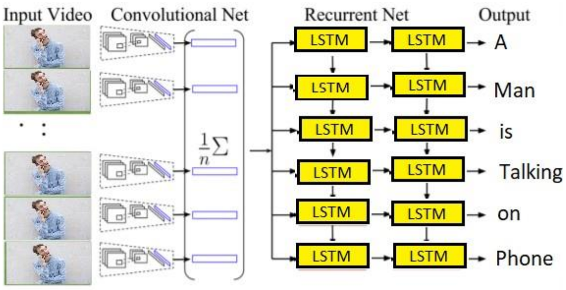

Figure 5: Mean-pool architecture [5]

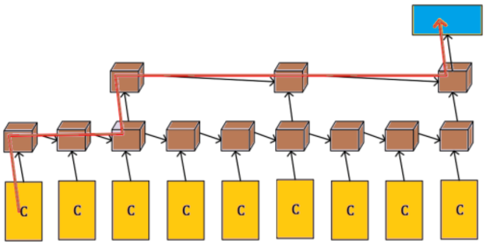

Figure 6: HRNE Model [1]

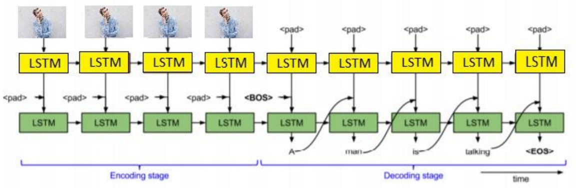

Figure 7: S2VT Model [6]

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

where v_i - output of CNN encoder. Capability of exploiting and recognizing the temporal dependencies between frames because all the arrays are treated with the equal essential and mean them up like this. In Encoder-decoder approach l frames are encodes first,one at a time to the LSTM first layer of two layer, where variable length is l.Natural language sentence is decoded from that latent representation with one word at a time, one time step output is feeded to the LSTM second layer in the mean time.

Figure 6 shows hierarchy in different stages.Complex temporal structure of video can be effectively by this model.In tasks of video captioning this achieve higher captioning scores (METEOR, BLEU).

It is not like stacking several layers of LSTM, the input is $(x_1,x_2,..x_t)$ separated into various chunks $(x_1,x_2,..x_n)$ $(x_{1+s},x_{2+s},..x_{n+s})$ $(x_{t-n+1},x_{t-n+2},..x_t)$ where s is the distance between two adjacent chunk Then output of one of the layer(1st) is passed as input to another layer. This come out to be successful method.

4.2 Soft Attention

Video is encoded with by mean of pixels or feature in all rills present in video.Then the output of this is passed into Image-Net pretrained CNN. But only limitation is that it has decrease the strength of learning temporal structure.To overcome this problem we use soft attention mechanisms, where one input is processed at time.In Soft attention we combine weights together of all frame level features,where word decoder affect weights. Un-normalized relevance of i^{th} temporal feature at decoder time step t and sequence of feature vector $V=(v_1,v_2,..v_n)$ is calculated as:

$$e^t_i = w^t \tanh(W_a h_{t-1} + U_a v_i + b_a) [6]$$

where h_{t-1} - invisible state of the decoder at previous time [10], v_i - vector representation of frame features of i^{th} frame, w, W_a, U_a, b_a - learned parameters.

After calculating relevance score, dynamic weight $a^{(t)}_i$ are calculated as:

$$a^{(t)}_i = \frac{exp(e^{(t)}_i)}{\sum_{j=1}^{n} exp(e^{(t)}_j)}$$

And then weight sum of temporal feature vector is computed as:

$$\Phi_t(V) = \sum_{i=1}^{n} a^{(t)}_i v_i [5]$$

5 METHODOLOGY

5.1 Gaussian Attention(GA)

Gaussian Attention can be defined as a technique to remove limitaions of the generic soft attention modals.The Gaussian Attention mainly aims to replace the Soft Attention technique. GA also have many advantages over SA, some of them are like fewer parameters and better captioning performance. To obtain the relevance score, we will model the input sequence with Gaussian Distribution. The decoder filter's the encoder sequence at each time. The input sequence is weighs by the GA on the basis of temporal location and the average and standard deviation model the shape of the distribution. To calculate a discontinuous free relevancy score e^y across the entire sequence of input, we adopts a function:

$$e^y = \sum_{i=1}^{N} \pi_i N(X|\mu^y_i, \Sigma^y_i) [10]$$

where N - number of Gaussian's, π_i - mixing coefficient, μ^y_i - unique mean, Σ^y_i - co variance matrix.

5.2 Attention Steering

Normal Attention model uses the collections of weights and during the training of the model they learn. And at the time of testing, the attentions are guided by weight matrix and that's the reason for limiting of attention mechanism. Mostly the video captioning

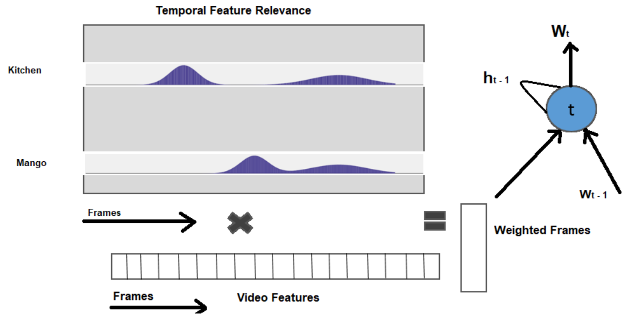

Figure 8: Distribution of Frame level features

data set possibly do not have a comprehensive represention of the activities and objects. Like "Eating of banana" is semantic action which is likely to appear at start of the video, then the trained model may show similar trend in the testing video. To get the more video oriented attention, we will make the web to watch the total video before going through the particular portion of the video. and we will add temporal concepts throught out the portion of the video. There are many ways to encode the details / summary of the video, but we will be using an LSTM network. LSTM is an architecture of artificial RNN and have the ability to remember the segments of the video but we have to made it retain both, the relevant frames and the context of the whole video. For the detailed attention, the latent representation of video is required but this restrictions of LSTM make it difficult. To tackle this situation we will use the temporal attention steering, which will guide the attention on the basis of the features of the test video. We did some research on the use of word label embeddings of the objects. We will we using Image-Net classifiers. Large no. of objects should be represented in the wild videos. We will we using a bottom-up grouping technique to deal with the over specifics class problems. A Phrase describes the object and whole scene of the context. In multi-object videos, identifying the individual object from the scenes is very challenging task. To tackle this challenge we will be using Edge Box. This will help us in getting the bounding box region of each frame. We will be computing the glove word embedding of Image-Net CNN class. The character embedded vectors is a Average Pooling to get a frame tag depiction. We came across the fact that Average Pooling class tag embeddings are wealthy in the meaning of words and sentence information's and are very close veracity sentences. Using the embedding word technique also reduces the dimensionality of features and hence resulting in reduced no. of graspable parameters. Frame CNN features can also be used directly instead of temporal word.

5.3 Video-2-vec representation

After the steering procedure, embedded word vector depiction of the video will be given as input to the captioning model. Word concepts and video level feature are important features. The Video2vec transformation is done through an embedding function $f : V \ S_v \longrightarrow$This embedding function maps the video with the frames.

5.3.1 Video2vec Activity. We will be using Activity-Net Classification dataset, as it will help the model to learn action and motion concepts. This dataset basically cover the broad range of human's complex activities. This dataset contain approx. 900 video hours spread across 300 activity classes. The videos which are labeled will be helpful in training a normal video based activity classifier. These videos are excellent for transfer learning feature for Microsoft Research video description Corpus and Microsoft Research Video to Text dataset. All videos in the dataset are collected from different websites which shares their video. We will be training 2 models, first model using RGB and second model is Optical flow inputs. We will be using Video-2-vec activity as feature before the loss layer and then the last connected layer is fine tuned during its making.

5.3.2 Gaussian Attention Hierarchy. We will integrate the proposed neural encoder with the Gaussian Attention. This will help the recurrent layer add more non-linearity to the Gaussian Attention model. To make the easy back propagation of loss, we will step the invisible state of LSTM in l-l layer to input to layer l. To distinguish between sequence of short clips of videos, the first layer is trained to learn temporal dynamics (Local), which are in between the short clips, and then the second layer is trained. The output of the entire video is in form of vector, which is also the output of the second layer.

5.4 Shot Boundary detection

From the CNN model, the feature which we extract will help in shot change spotting between two shots, in rills of videos. The cosine distance (DELTA), between two consecutive CNN feature vector frames a_i & a_y can be defined

$$\Delta(x, y) = \cos(a_x, a_y \quad) = \frac{a_x . a_y}{|a_x|.|a_y|}$$

where Δ ranges from (0, 1), higher value means higher probability of a boundary cut. Cosine distance do not need additional steps to normalize unlike Euclidean Distance.

5.5 Multi-stream Hierarchical Boundary Model

The video stream will be given as input to the encoding stage and the local features (x_1, x_2, x_3, ... x_n) will be given to first layer. These two layers then provide us a output as a 2 vector sequence. First one is equally spaced [w_1, w_2, w_3 w_p]. and second one is clip levels [z_1, z_2, z_3, z_q] The first equally spaced vector sequence get M output. These M outputs were given by the first layer, where M = n / k, n - feature count(input), k - designed stride value. the second clipped level vector sequence make use of the information on the short boundary, which will be lead by a learned vector. This learned vector is based on cosine distance:

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

where W_{yd} - learned weights , b_{yd} - learned bias, y_i - output of first layer of each time step. Figure 8, show that how video will be encoded by combining the equally spaces and clipped level vector characteristic. To provide input to caption decoder, we will provide a blend of frame level, equally spaced sequence and clipped level(Detected Boundaries) vector sequence. The model will modify the boundary weights, at each time step, to draw out the details

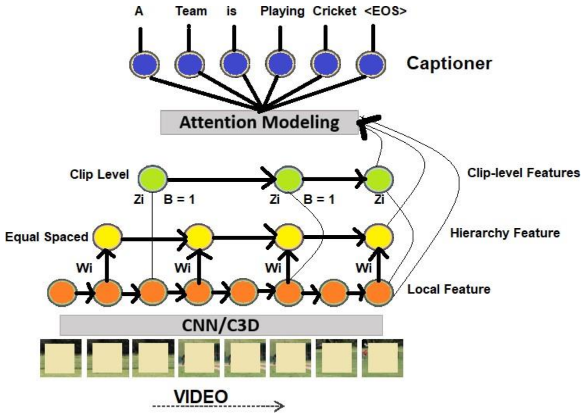

Figure 9: MSHB Model [10]

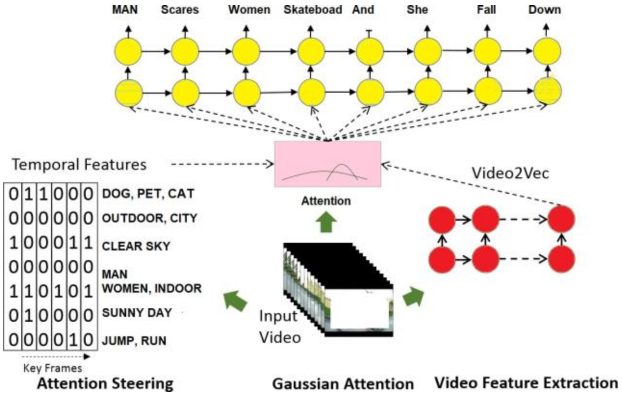

Figure 10: Video Captioning model using Gaussian Steering[10]

from the parts of the video, this will very efficient in encoding the videos which are complex in nature.

5.6 Gaussian steering model

Attention steering, Video2vec encoder and Gaussian Attention are the main components of our video captioning framework. The input to sentence generation engine is given from all the 3 main components to generate the sequence of meaning words. For the generation of sequence like natural language sentence RNN is the natural choice. There are certain limitations of RNN such as vanishing and exploding gradient problem, to overcome this limitation we will be using LSTM variant of RNN, this will help in generating the sentence by learning, by using short temporal and long temporal technique.

6 CONCLUSION

In recent years video caption generation techniques has made advancements due to works in various deep-learning techniques and has proved to a milestone in the accuracy of video captioning. Image retrieval efficiency which is based on content can be improved by text description of the video which is divided into images. In this paper a hierarchical framework is introduced for getting into the video and video attributes and length of video are used as features to get the attention. Also, in this paper multi-stream captioning us being described that can handle simple as well as complex videos and generate their captions This project has further scopes expanding to various applications in security, surveillance, military, medicine and for visually impaired persons etc.

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.

[7] Yu, Y., et al., *Video Captioning and Retrieval Models with Semantic Attention.* arXiv preprint arXiv:1610.02947, 2016.

[8] Caba Heilbron, F., et al, *Activitynet: A large-scale video benchmark for human activity understanding. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[9] Xu, J., L. Song, and R. Xie, *Shot boundary detection using convolutional neural networks. in Visual Communications and Image Processing (VCIP), 2016.* 2016. IEEE.

[10] Nguyen, T. H. (2017). *Automatic Video Captioning using Deep Neural Network.*

[11] Farzana, U. A., Abirami, S., Srivani, M. (2019, December). A Framework For Captioning The Human Interactions. In *2019 11th International Conference on Advanced Computing (ICoAC)* (pp. 13-17). IEEE.

[12] Amirian, S., Rasheed, K., Taha, T. R., Arabnia, H. R. (2020). Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap. *IEEE Access*, *8*, 218386-218400.

[13] Yang, Y., Zhou, J., Ai, J., Bin, Y., Hanjalic, A., Shen, H. T., Ji, Y. (2018). Video captioning by adversarial LSTM. *IEEE Transactions on Image Processing, 27(11),* 5600-5611.

[14] Lee, S., Kim, I. (2018). Multimodal feature learning for video captioning. *Mathe- matical Problems in Engineering, 2018.*

[15] Islam, S., Dash, A., Seum, A., Raj, A. H., Hossain, T., Shah, F. M. (2021). Explor- ing Video Captioning Techniques: A Comprehensive Survey on Deep Learning Methods. *SN Computer Science*, 2(2), 1-28.

[16] Li, S., Tao, Z., Li, K., Fu, Y. (2019). Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4), 297-312.

[17] Liu, S., Bai, L., Hu, Y., Wang, H. (2018). Image Captioning Based on Deep Neural Networks. In *MATEC Web of Conferences* (Vol. 232, p. 01052). EDP Sciences.

[18] Mathur, P., Gill, A., Yadav, A., Mishra, A., Bansode, N. K. (2017, June). Cam- era2Caption: a real-time image caption generator. In *2017 International Conference on Computational Intelligence in Data Science (ICCIDS)* (pp. 1-6). IEEE.

[19] Sharma, G., Kalena, P., Malde, N., Nair, A., Parkar, S. (2019, April). Visual image caption generator using deep learning. In *2nd International Conference on Advances in Science Technology (ICAST).*

[20] Gao, L., Guo, Z., Zhang, H., Xu, X., Shen, H. T. (2017). Video captioning with attention-based LSTM and semantic consistency. *IEEE Transactions on Multimedia, 19(9),* 2045-2055.

[21] Song, J., Guo, Y., Gao, L., Li, X., Hanjalic, A., Shen, H. T. (2018). From deterministic to generative: Multimodal stochastic RNNs for video captioning. *IEEE transactions on neural networks and learning systems, 30(10),* 3047-3058.

[22] Oura, S., Matsukawa, T., Suzuki, E. (2018, July). Multimodal Deep Neural Network with Image Sequence Features for Video Captioning. In *2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7).* IEEE.

[23] Amaresh, M., Chitrakala, S. (2019, April). Video captioning using deep learning: An overview of methods, datasets and metrics. In *2019 International Conference on Communication and Signal Processing (ICCSP) (pp. 0656-0661).* IEEE.

[24] Wang, C. Y., Liaw, P. S., Liang, K. W., Wang, J. C., Chang, P. C. (2019, September). Video Captioning Based on Joint Image–audio Deep Learning Techniques. In *2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin)* (pp. 127-131). IEEE.

[25] Sehgal, S., Sharma, J., Chaudhary, N. (2020, June). Generating Image Captions based on Deep Learning and Natural language Processing. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 165-169).* IEEE.

[26] Goehring, T., Bolner, F., Monaghan, J. J., Van Dijk, B., Zarowski, A., Bleeck, S. (2017). Speech enhancement based on neural networks improves speech intelligi- bility in noise for cochlear implant users. *Hearing research, 344, 183-194.*

[27] Andra, M. B., Usagawa, T. (2021). *Improved Transcription and Speaker Identifi- cation System for Concurrent Speech in Bahasa Indonesia Using Recurrent Neural Network.* IEEE Access.

[28] Khalil, K., Dey, B., Kumar, A., Bayoumi, M. (2021, May). A Reversible-Logic based Architecture for Long Short-Term Memory (LSTM) Network. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 1-5).* IEEE.

[29] Li, T., Hua, M., Wu, X. (2020). A hybrid CNN-LSTM model for forecasting particulate matter (PM2. 5). *IEEE Access, 8, 26933-26940.*

[30] Liu, Y., Gong, C., Yang, L., Chen, Y. (2020). DSTP-RNN: A dual-stage two-phase attention-based recurrent neural network for long-term and multivariate time series prediction. *Expert Systems with Applications,* 143, 113082.