ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2695

Generating Video Descriptions with Attention-Driven LSTM Models in Hindi Language

Naman*[1]; Harsh Nagar[2]; Dhruv[3]; Vansh Gupta[4]
Eelectronics & Communication (AI & ML), Netaji Subhas University of Technology,
Dwarka, New Delhi, India – 110078

Abstract:- This research addresses the existing gap in video descriptions for regional languages, with a particular emphasis on Hindi. Motivated by a thorough review of available literature, it was observed that languages like Hindi are inadequately represented in this domain. Consequently, we initiated the project titled "Generating Video Descriptions with Attention-Driven LSTM Models in Hindi Language" to enhance accessibility and inclusion of Hindi multimedia content. Leveraging advanced LSTM models and utilizing the VATEX dataset, our objective is to pioneer advancements in regional narrative video production. By venturing into unexplored terrain, we not only contribute to the promotion of Indian language and culture but also establish a precedent for exploring narrative films in other regional languages. This research is strategically designed to foster diversity, integration, and propel broader advancements at the intersection of natural language processing and multitasking. Our findings demonstrate that our approach yields competitive performance when compared to state-of-the-art video captioning baselines such as BLEU and METEOR. This signifies the efficacy of our methodology in enhancing the quality of video descriptions, thereby contributing significantly to the field of regional language video captioning.

Keywords:- *Video Description, Attention-Based LSTM, VATEX, Hindi Language.*

I. INTRODUCTION

Video description has gained attention due to advances in computer vision, NLP, and machine learning. It involves describing video content with natural language sentences, serving various purposes like human-robot interaction and accessibility for the visually impaired.

Two main approaches exist: template-based language models and sequence learning. Template models use predefined templates to structure sentences, ensuring consistent syntax; sequence learning directly translates video content into phrases by extracting features and generating subtitles based on data. An encoder-decoder network offers an efficient solution. The encoder processes the input video, producing a fixed-dimensional vector fed into the decoder, which generates words sequentially.

In the early 2010s, the field of video captioning and description generation witnessed a significant evolution, driven by advancements in computer vision and natural language processing. This transformative shift began with the introduction of large-scale datasets and neural network-based models. Key papers such as "Show and Tell: A Neural Image Caption Generator" by Vinyals et al. in 2015 played a pivotal role in pioneering the fusion of image analysis and language generation, setting the stage for modern video captioning. This progression from basic audio captioning in the 1990s to the sophisticated video captioning of the 2010s highlights the rapid growth of technology in enhancing media accessibility and usability. This literature survey explores the key developments that have shaped this dynamic field.

Captioning is the process of adding text to visual content, providing information about the content for various purposes. There are several types of captioning, each serving different needs and audiences. Here are some common types of captioning: closed captions, open captions, subtitles, live captions, *descriptive video service (DVS) or audio descriptions. There are sone major applications of videos captioning, i.e. security surveillance, improved indexing for search engine optimizer (SEO), text-to-speech of captions for visually impaired.

The landscape of video and image captioning is evolving with remarkable innovations highlighted in recent studies. The "Attention-based Densely Connected LSTM for Video Captioning" [2] underscores the pivotal role of contextual integration via Dense LSTM for enriched narrative construction in video captioning. In parallel, "Dense Video Captioning Based on Local Attention" [3] introduces a paradigm shift towards Dense Video Captioning Locally (DVCL), focusing on global feature dependence and enhancing word-frame correlation for precision. The realm of linguistic diversity in captioning is expanded through "Attention-Based Hybrid Deep Learning for Bengali Video Captioning" [4], setting new benchmarks in Bengali language video captioning with a hybrid deep learning approach. Similarly, "Hybrid Architecture using CNN and LSTM for Image Captioning in Hindi Language" [5] leverages the synergy between CNN and LSTM to achieve significant improvements in Hindi image captioning. The "pLSTM Framework for Video Captioning in VATEX2020" [6] marks a milestone with its pioneering framework, establishing a new benchmark in the VATEX2020 challenge. A broad perspective on the field's evolution and future directions is provided by "Video Captioning: A Comparative Review of Progress and Opportunities" [7], detailing the progress and identifying potential opportunities in video captioning. The "Attention-Based Dual Learning Approach for Video Captioning" [8] delves into a dual-learning paradigm, enriching the video captioning process by synergizing insights from both videos and generated captions. Lastly, "Video Captioning with Attention-based LSTM and Semantic Consistency" [9] enhances captioning accuracy by marrying attention mechanisms with LSTM, underscored by a focus on semantic consistency, showcasing the cumulative advancements and diverse methodologies propelling the field forward.

II. PREVIOUS RELATED WORK

A. Convolution Neural Network (CNN)

One type of deep, feed-forward artificial neural network is the convolutional neural network (CNN). CNNs are one of the most popular algorithms in deep learning. They are capable of achieving state-of-the-art results and have a wide range of applications in image recognition, speech recognition, and natural language processing. Convolutional Neural Networks (CNNs) have been widely used in image recognition tasks like object detection, classification, face recognition, and other computer vision problems. In CNN's, we are not required to fix the length k of a feature map as a given value; instead, we can vary it for different layers and then weight each feature map by its receptive field size to define a feature map size of the layer. The Convolution Operation extracts the input image's high-level features, such as edges. It is unnecessary to limit ConvNets to a single Convolutional Layer. Typically, low-level features like edges, colour, gradient direction, and so on are captured by the first ConvLayer. The architecture adjusts to the high-level properties by adding layers, giving us a network that understands the photos in the dataset in a way that is similar to how we would.

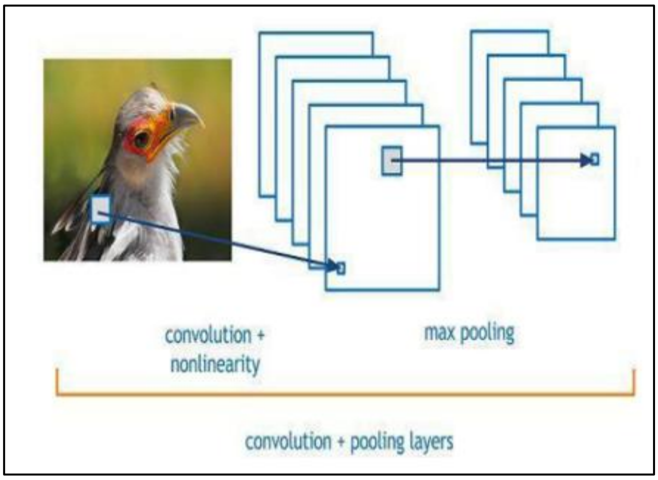


Fig 1: Convolution Layers

B. VGG-16

VGGNet-16 is made up of 16 convolutional layers and is particularly appealing due to its consistency. It is the most popular method for extracting features from photos at the moment. VGG can be achieved using transfer learning, in which the model is pre-trained on a dataset and the parameters are changed for improved precision. The ImageNet dataset contains 14 million photos divided into 1000 classes, and this model has a top-5 test precision of 92.7 percent. VGG-16 was one of the top-performing architectures in the 2014 ILSVRC competition. With a 7.32 percent top-5 classification error, it came in second in the classification task (only behind GoogLeNet with a classification error of 6.66 percent). It also won the position error localization task with a score of 25.32 percent.

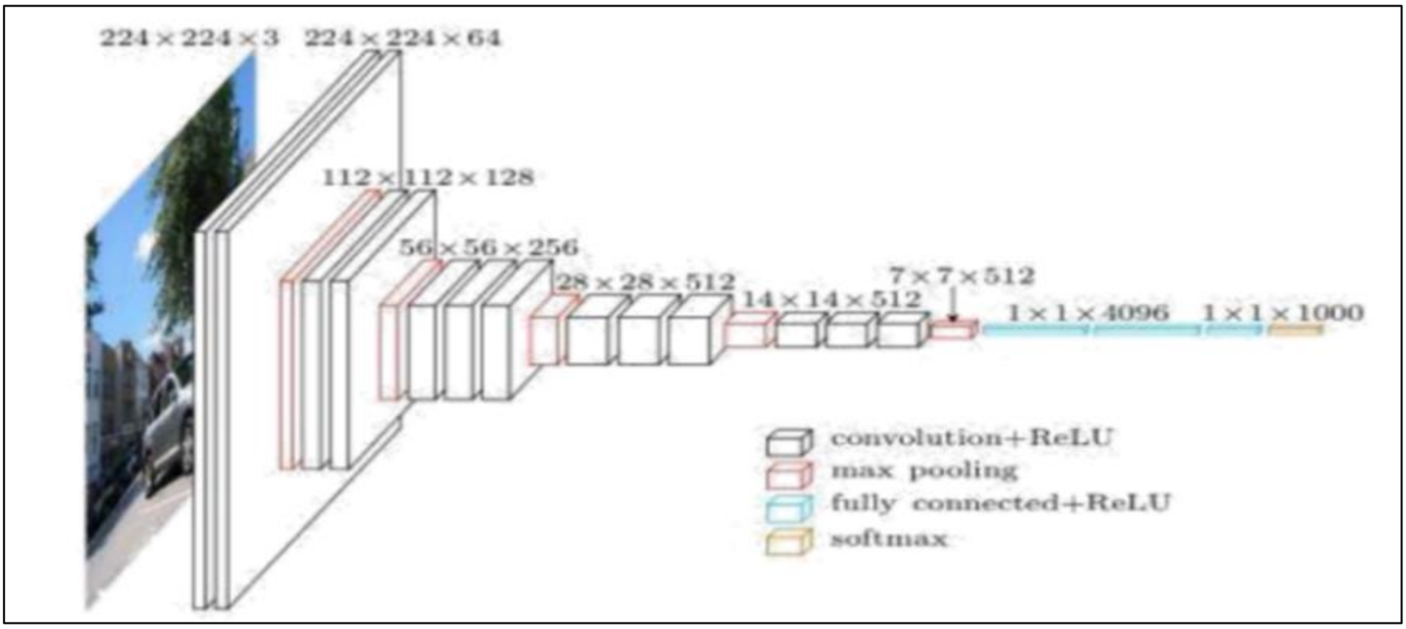


Fig 2: VGG-16 Overview

C. XCEPTION

The Xception Model is divided into three sections: Entry Flow, Middle Flow and Exit Flow.

Two sets of convolution layers follow the ReLU layer in the Entry Block. Separate convolution layers, pooling layers, and skip connections are also included. The figure shows the filter size and stride of each layer. If we have knowledge from prior layers that may be relevant in subsequent layers, we employ skip connections. Similarly, the middle block comprises three sets of ReLU, followed by eight repetitions of Separable Convolution. The figure depicts the input and output shapes. There is a worldwide average pooling of the exit flow (GAP). This layer replaces completely connected layers with GAP to decrease parameters. It reduces the output to 1x1xc for the input matrix of m x n x c. For each channel, it takes an average across the entire 2D matrix. The matrix is then flattened into a single vector and fed to a fully connected layer, which finally feeds it to the final layer, which uses logistic regression to identify classes. The model was trained on 1000 ImageNet Dataset classes.

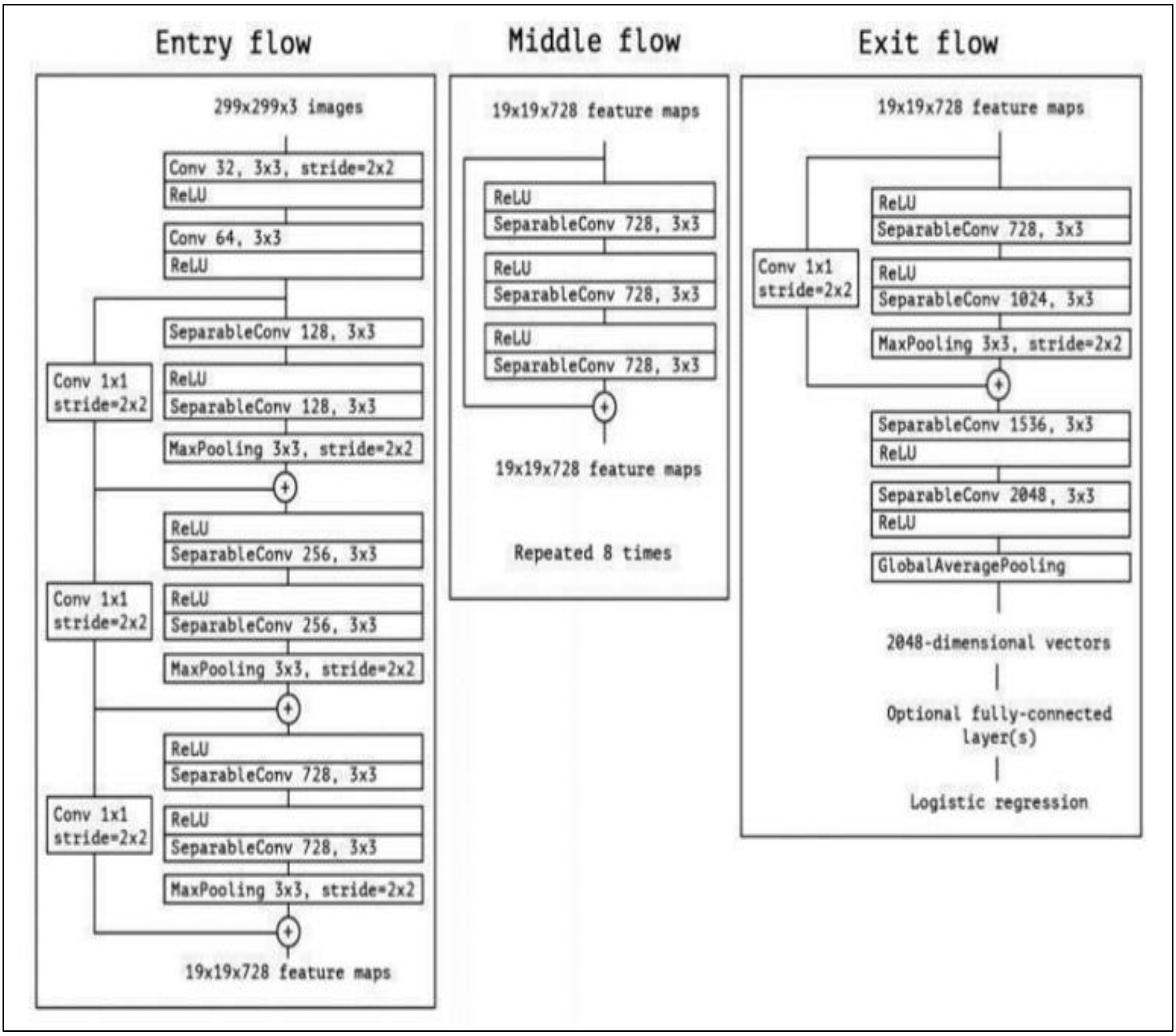


Fig 3: XCEPTION net

D. Recurrent Neural Network (RNN)

A RNN is a neural network in which the current step's output is used as an input for the following phase. It allows for a more complicated temporal activity. In typical neural networks, all inputs and outputs are independent of one another, but in circumstances like sequence generation, the prior words are necessary to predict the next word of a phrase, and so the previous words must be remembered. RNN was born, and with the help of a Hidden Layer, it was able to address the problem. There are five different types of RNNs.

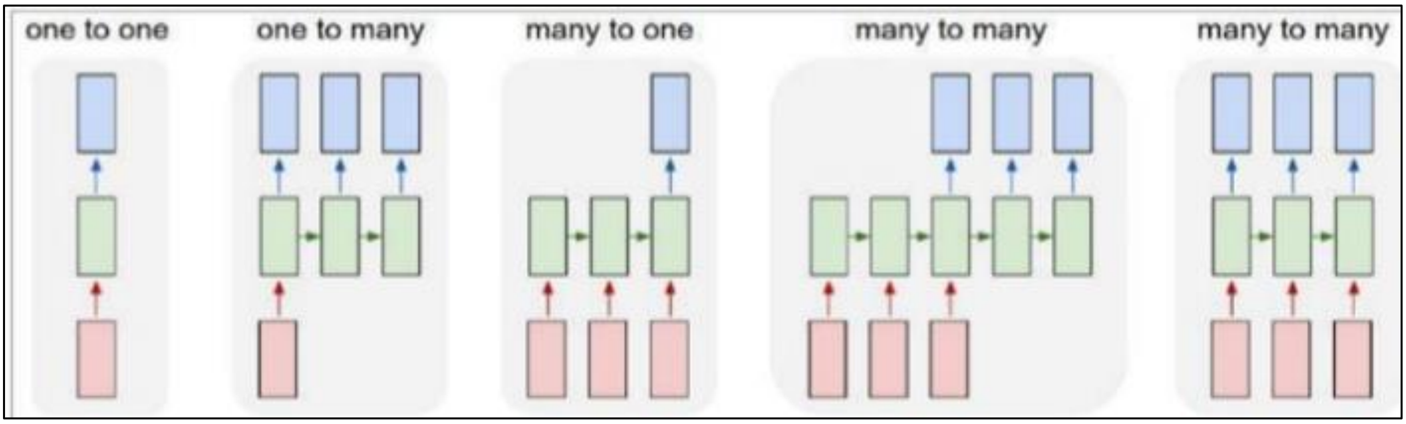


Fig 4: Types of RNNs

E. LSTM

RNN assists in connecting earlier information to current activities, such as understanding the context of the current sentence by using previous sentences. To complete a task, we may simply require recent knowledge. For example, while creating sentences, we sometimes only need a few words to generate the following word. For example, in the sentence "I awake at 7 a.m. and drink banana shake," the word that comes after is "juice." All it takes is a phrase beginning with "drink banana" to predict the word "shake." When the relevant information and the place where it is needed are both tiny, RNN can learn to use the existing information. In other cases, further context is needed to decide which is the last word. For instance, "I can write in and live in India." "English," "Hindi," or any of the other few possible words may be the one that comes next, but "German" or "Spanish" are improbable choices. In this case, the context "I live in India" is necessary. It's possible that the distance between the essential information and the location where it's needed grows significantly. RNN is unable to learn to connect the information as the gap widens. The issue with RNN is this. They are incapable of dealing with long-term reliance. Short-term memory is another name for this.

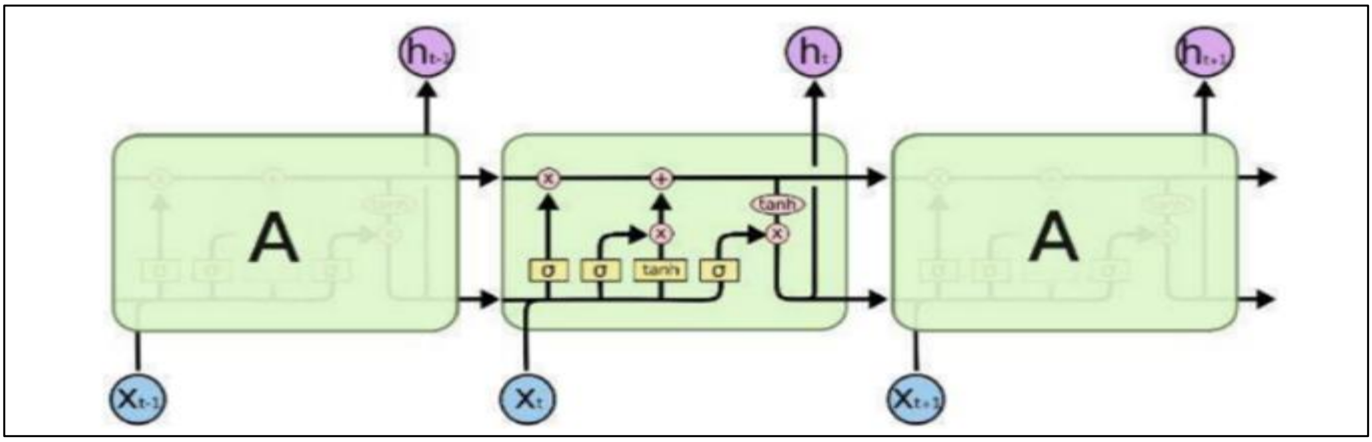


Fig 5: LSTM Unit

III. PROPOSED APPROACH & EXPERIMENNT

We were motivated to launch the project "Generating Video Descriptions with Attention-Driven LSTM Models in Hindi." We observed a dearth of research in regional languages like Hindi, especially in the realm of video description generation. Recognizing this uncharted territory, we aim to leverage advanced LSTM models and VATEX dataset to enhance accessibility and inclusivity in Hindi multimedia content. Our project not only contributes to the Hindi language but also sets a precedent for the exploration of other regional languages, ultimately fostering diversity and inclusivity in the multimedia landscape.

A. Dataset

- ➢ *VATEX - A Large-Scale, High-Quality Multilingual Dataset for Video-and Language Research [1]*

A brand-new, extensive multilingual video description dataset called VATEX has 825,000 captions in both Chinese and English for over 41,250 videos. The following are the main distinctive characteristics of VATEX. Firstly, it has descriptions in both Chinese and English at scale, which can help numerous multilingual research that are limited by datasets that are only available in one language. Second, VATEX has the highest number of clip-sentence pairs in the corpus. Each video clip has numerous distinct phrases attached to it, and each caption is distinct. Thirdly, with 600 total human activities covered, VATEX offers more thorough but representative video footage.

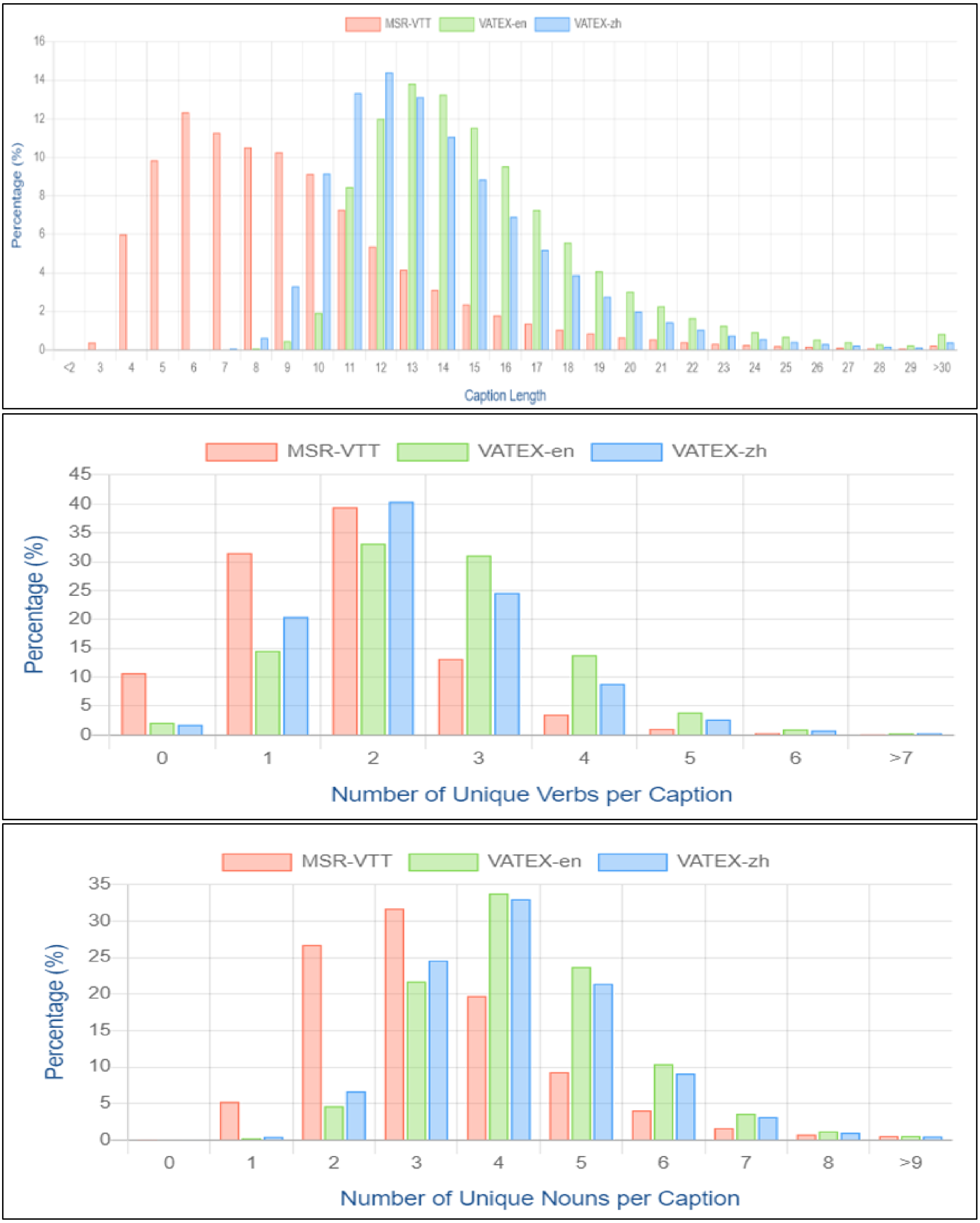
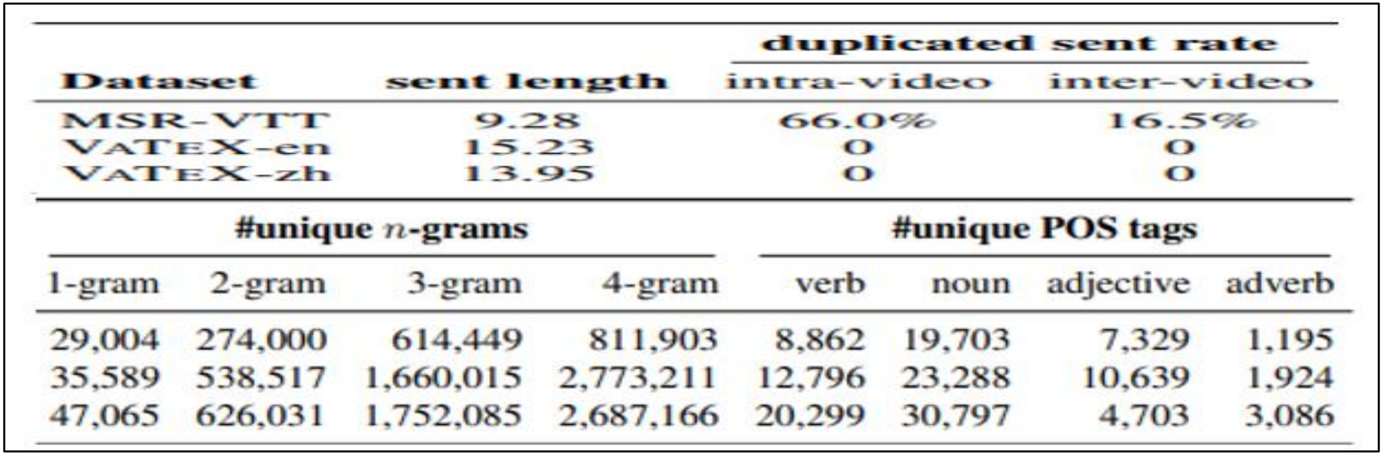


Fig 6: Statistical Histogram Distributions on MSR-VTT, VATEX-en, and VATEX-zh. Compared to MSR-VTT, the VATEX Dataset Contains Longer Captions, Each with More Unique Nouns and Verbs

B. *Dataset Analysis*

Table 1: Average Data



Dataset	sent length	duplicated sent rate	
		intra-video	inter-video
MSR-VTT	9.28	66.0%	16.5%
VATEX-en	15.23	0	0
VATEX-zh	13.95	0	0

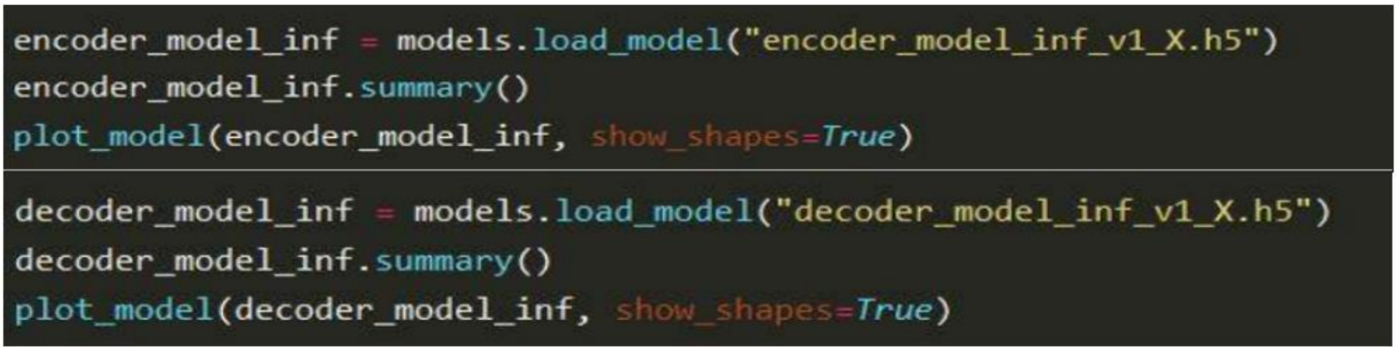
#unique n-grams				#unique POS tags			
1-gram	2-gram	3-gram	4-gram	verb	noun	adjective	adverb
29,004	274,000	614,449	811,903	8,862	19,703	7,329	1,195
35,589	538,517	1,660,015	2,773,211	12,796	23,288	10,639	1,924
47,065	626,031	1,752,085	2,687,166	20,299	30,797	4,703	3,086

In Table 1, We do a quick comparison of the overall statistics of the available datasets for video descriptions. We perform a thorough examination in this part between our VATEX dataset and the MSR-VTT dataset, which is the most similar to VATEX in terms of scale and domain and is the most used benchmark for video captioning. We divided VATEX into the English corpus (VATEX-en) and the Chinese corpus (VATEX-zh) for comparison because MSR-VTT only contains an English corpus. VATEX contains 413k English and 413k Chinese captions explaining 41.3k unique videos from 600 activities, whereas MSR-VTT has 200k captions explaining 7k films from 257 activities. In comparison to MSR-VTT, the subtitles in VATEX-en and VATEX-zh are larger, longer, and more comprehensive. VATEX-en, VATEX-zh, and MSR-VTT had average caption lengths of 15.23, 13.95, and 9.28, respectively. We compare the unique n-grams and part-of-speech (POS) tags (such as verb, noun, adverb, etc.) between MSR-VTT, VATEX-en, and VATEX-zh in order to evaluate the linguistic complexity. This comparison highlights the advantages of VATEX over MSR-VTT as well as the distinctions between the English and Chinese corpora. Our VATEX datasets encompass a greater range of activities, objects, and visual settings, and thus reflect a bigger diversity of caption styles.

We also carry out detailed analyses of caption diversity comparisons. First, as Table 1 illustrates, MSR-VTT suffers from a serious duplication problem because 66.0% of the movies have some identical duplicate captions. In contrast, our VATEX datasets are free from this issue and ensure that each caption inside a single video is distinct. Our VATEX datasets' captions show more variability both inside and throughout the corpus, indicating that our VATEX may be used as an improved benchmark for video retrieval. Next, as a more intuitive measure of the lexical richness and caption diversity, we propose the Type-Caption Curve, which is derived from the type-token vocabulary curve but particularly designed for the caption corpora here. The total number of captions and the number of distinct vocabulary items (types) are computed for each corpus. We therefore plot the number of sorts vs the number of captions for MSR-VTT, VATEX-en, and VATEX-zh (see Figure 4, where we use 4-grams as the types). These type-caption curves show how our VATEX datasets are more linguistically diverse and sophisticated, with information on lexical competency (vocabulary size) and lexical style or caption variation (vocabulary usage) discernible via inference.

C. *Implementation*

We've already trained the model (encoder and decoder) and stored it as a.h5 file for future usage. An encoder model and a decoder model make up the overall model. As a result, we must first load these two models:



```
encoder_model_inf = models.load_model("encoder_model_inf_v1_X.h5")
encoder_model_inf.summary()
plot_model(encoder_model_inf, show_shapes=True)
```

```
decoder_model_inf = models.load_model("decoder_model_inf_v1_X.h5")
decoder_model_inf.summary()
plot_model(decoder_model_inf, show_shapes=True)
```

Fig 7: Loading the Model (Encoder + Decoder)

Once these two models have been loaded successfully, we can use them to generate captions for the given videos. We have a generate () function that receives an input sequence as an argument and returns the predicted sentence. Navigate to the root directory of the project using windows cmd/ PowerShell / terminal.

D. Results Obtained

Among the outcomes we have attained are:

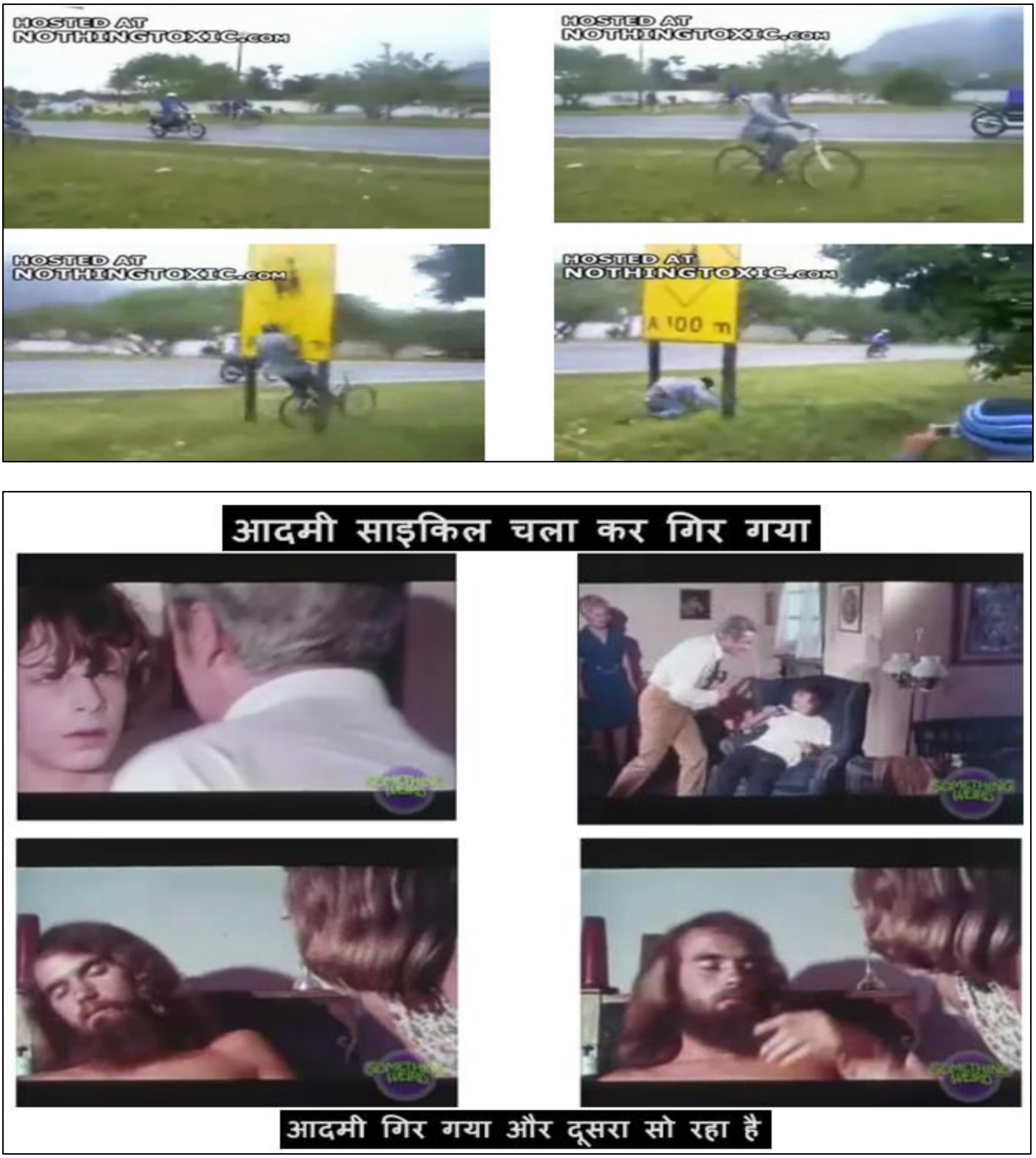

HOSTED AT NOTHINGTOXIC.com
A 100 m
आदमी साइकिल चला कर गिर गया
आदमी गिर गया और दूसरा सो रहा है

Fig 8: Output 1 & 2

E. Performance Analysis & BLEU Score

BLEU stands for Bilingual evaluation understudy. It is used to evaluate text output. Table 6.5.1 shows how Bleu Score is interpreted and based on the Bleu score; we analyze the performance of the model. It can be observed that we desire a higher bleu score. Our model does not account for grammatical errors but the gist of the caption so our desired Bleu score is any number greater than 30.

Table 2: Bleu Score Interpretation

BLEU Score	Interpretation
<10	Almost useless
10-19	Hard to get the gist
20-29	The gist is clear but has grammatical error
30-39	Understandable to good translation
40-49	High quality translation
50-59	Very high quality and fluent translation
≥60	Quality often better than human

Bleu Score is calculated using the following formulas:
N-Gram Precision

$$p_n = \frac{\sum_{n-gram\, \epsilon\, hyp} count_{clip}(n-gram)}{\sum_{n-gram\, \epsilon hyp} count\,(n-gram)}$$

$$Bleu = Be^{\frac{\sum_{n=1}^{N} p_n}{N}}$$

N grams refer to N words combinations next to each other in a sentence. In unigram precision, we take individual words. Brevity penalty is used for shorter sentences than the reference. Our model resulted in a Bleu score of 24 which is significantly good for a simpler model and smaller dataset. This means that our model can understand the content of the video but has some grammatical errors. Since our model does not account for grammatical errors, our model will evaluate to a low Bleu score if we consider bigram or trigram precision. Also, Bleu score does not take into consideration the synonyms of words.

IV. CONCLUSION

The findings of our research underscore the effectiveness of our methodology, as evidenced by competitive or superior performance compared to state-of-the-art video captioning baselines such as BLEU and METEOR. This validation signifies a significant stride forward in enhancing the quality of video descriptions for regional languages, particularly Hindi, thereby making a noteworthy contribution to the field of regional language video captioning. We bring our research journey to a culmination by summarizing key findings and insights in the conclusion. Reflecting on the implications of our study, we provide a cohesive understanding of the contributions made and the significance of our research. In essence, our research not only fills a critical void but also establishes a framework for future exploration in the domain of regional language video production. As we move forward, we anticipate that our findings will serve as a catalyst for continued innovation and improvement in the realm of natural language processing and multimedia accessibility, ultimately fostering a richer and more inclusive digital landscape. It can be observed that we desire a higher bleu score. Our model does not account for grammatical errors but the gist of the caption so our desired Bleu score is any number greater than 30. Our model resulted in a Bleu score of 24.

ACKNOWLEDGEMENT

We express sincere gratitude to our esteemed academic guide and mentor, Prof. RK Sharma, whose continuous guidance and unwavering support have been pivotal throughout our B.Tech. journey at Netaji Subhas University of Technology (NSUT), influencing not only this research but also various other course works. Special appreciation is extended to the dedicated professors and encouraging friends at NSUT, whose support has significantly contributed to the success of our project. Last but not least, our heartfelt thanks go to our parents and siblings for their boundless love and steadfast support. Their encouragement has been the driving force that made the completion of this work possible.

REFERENCES

[1]. Xin Wang, Jiawei Wu, Junkun Chen, Lei Li2=, Yuan-Fang Wang, William Yang Wang (2020) VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research, University of California, Santa Barbara, CA, USA, Byte Dance AI Lab, Beijing, China, arXiv:1904.03493v3.

[2]. Yongqing Zhu, Shuqiang Jiang (2019) Attention-based Densely Connected LSTM for Video Captioning, Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, 100190, China University of Chinese Academy of Sciences, Beijing, 100049, China, MM '19, October 21–25, 2019, Nice, France.

[3]. Yong Qian, Yingchi Mao, Zhihao Chen, Chang Li, Olano Teah Bloh, Qian Huang (2023) Dense video captioning based on local attention, Key Research and Development Program of China, Grant/Award Number: 2022YFC3005401; Key Research and Development Program of Yunnan Province, Grant/Award Numbers: 202203AA080009, 202202AF080003; the Key Technology Project of China Huaneng Group, Grant/Award Number: HNKJ20-H46, DOI: 10.1049/ipr2.12819.

[4]. Md. Shahir Zaoad, M.M. Rushadul Mannan, Angshu Bikash Mandol, Mostafizur Rahman, Md. Adnanul Islam, Md. Mahbubur Rahman (2023) An attention-based hybrid deep learning approach for Bengali video captioning, Department of Computer Science and Engineering, Military Institute of Science and Technology, Dhaka 1216, Bangladesh.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2695

[5]. Ayush Kumar Poddara, Dr. Rajneesh Rani (2023) Hybrid Architecture using CNN and LSTM for Image Captioning in Hindi Language, Dr B R Ambedkar National Institute of Technology, Jalandhar, Punjab, India, Peer-review under responsibility of the scientific committee of the International Conference on Machine Learning and Data Engineering 10.1016/j.procs.2023.01.049.

[6]. Alok Singh, Salam Michael Singha, Loitongbam Sanayai Meetei, Ringki Das, Thoudam Doren Singh, Sivaji Bandyopadhyay, (2023)] VATEX2020: pLSTM framework for video captioning, Department of Computer Science and Engineering, National Institute of Technology Silchar Assam, India, Center for Natural Language Processing, National Institute of Technology Silchar Assam, India.

[7]. Daniela Moctezuma, Tania Ram´ırez-delReal, Guillermo Ruiz, Oth´on Gonz´alezCh´avez1 (2022) Video Captioning: a comparative review of where we are and 59 which could be the route, Centro de Investigaci´on en Ciencias de Informaci´on Geoespacial AC, Circuito Tecnopolo II , Aguascalientes, 20313, Mexico, Consejo Nacional de Ciencia y Tecnolog´ıa (CONACyT), Av. Insurgentes Sur 1582, Ciudad de Mexico, 03940, Mexico.

[8]. Wanting Ji a, Ruili Wang b, Yan Tian b, Xun Wang (2021) An attention based dual learning approach for video captioning, School of Information, Liaoning University, Shenyang, China, School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou, China.

[9]. Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu and Heng Tao Shen, Senior Member, IEEE (2017) Video Captioning with Attention-based LSTM and Semantic Consistency, School of Computer Science and Engineering, University of Electronic Science and Technology of China, 611731. Hanwang Zhang is with Department of Computer Science, Columbia University, USA. Heng Tao Shen is the correspondence author, Citation information: DOI 10.1109/TMM.2017.2729019, IEEE.

[10]. Olivastri, Silvio & Singh, Gurkirt & Cuzzolin, Fabio. (2019). End-to-End Video Captioning. 1474-1482. 10.1109/ICCVW.2019.00185.

[11]. Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video Description: A Survey of Methods, Datasets, and Evaluation Metrics. ACM Comput. Surv. 52, 6, Article 115 (January 2020), 37 pages. DOI: https://doi.org/10.1145/3355390.

[12]. Lee, Sujin & Kim, Incheol. (2018). Multimodal Feature Learning for Video Captioning. Mathematical Problems in Engineering. 2018. 1-8.

[13]. JX. Hua, X. Wang, T. Rui, F. Shao and D. Wang, "Adversarial Reinforcement Learning with Object-Scene Relational Graph for Video Captioning," in IEEE Transactions on Image Processing, vol. 31, pp. 2004-2016, 2022, doi: 10.1109/TIP.2022.3148868.

[14]. Iashin, Vladimir, and Rahtu, E. 2020. Multi-modal dense video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

[15]. J. Deng, L. Li, B. Zhang, S. Wang, Z. Zha and Q. Huang, "Syntax-Guided Hierarchical Attention Network for Video Captioning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 2, pp. 880-892, Feb. 2022, doi: 10.1109/TCSVT.2021.3063423.

[16]. S. Liu, Z. Ren and J. Yuan, "SibNet: Sibling Convolutional Encoder for Video Captioning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, v vol. 43, no. 9, pp. 3259-3272, 1 Sept. 2021, doi: 10.1109/TPAMI.2019.2940007.

[17]. Harsh Agrawal, Karan Desai, Xinlei Chen, Rishabh Jain, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. arXiv preprint arXiv:1812.08658, 2018.

[18]. Ozan Caglayan, Lo¨ıc Barrault, and Fethi Bougares. Multimodal attention for neural machine translation. arXiv preprint arXiv:1609.03976, 2016.

[19]. Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense- captioning events in videos. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), pages 706–715, 2017.

[20]. Gunnar A Sigurdsson, Gul Varol, Xiaolong Wang, Ali ¨ Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In Proceedings of the 2016 European Conference on Computer Vision (ECCV), pages 510–526, 2016.

[21]. Wajdi Zaghouani, Nizar Habash, Ossama Obeid, Behrang Mohit, Houda Bouamor, and Kemal Oflazer. Building an arabic machine translation post-edited corpus: Guidelines and annotation. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), pages 1869–1876, 2016.

[22]. Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus- based image description evaluation. Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4566–4575, 2015.

[23]. Xirong Li, Xiaoxu Wang, Chaoxi Xu, Weiyu Lan, Qijie Wei, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning and retrieval. IEEE Transactions on Multimedia, 2019.

[24]. Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[25]. Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Lo¨ıc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. In Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL), 2018.