AI Driven Approach for Description of Visual Contents

Radha Indoriya
Computer science and engineering
Chandigarh University
Mohali India
indoriya.radha@gmail.com

Pulkit Aggarwal
Computer science and engineering
Chandigarh University
Mohali India
pulkitaggarwal921@gmail.com

Chirag Jindal
Computer science and engineering
Chandigarh University
Mohali India
Jindalchirag586@gmail.com

Satyam Gupta
Computer science and engineering
Chandigarh University
Mohali India
Guptasatyam0704@gmail.com

Gurmehardeep Singh
Computer science and engineering
Chandigarh University
Mohali India
sgurmehardeep@gmail.com

***Abstract*—During the previous decade Deep Learning technology has entered numerous fields where it produces outputs comparable to or beyond human professional capabilities. The approaches demonstrate excellent performance thanks to their higher-level functionality together with access to advanced computing resources. This paper designs a cinematographic solution that crosses video captioning and Natural Language Processing to produce video titles together with abstracts. The method enables various applications including video database management and security monitoring in addition to search engines and the film sector. Our video description model uses deep learning since it obtains visual features by processing video frames with convolutional neural networks and distributes these CNN-generated inputs to a language model built with long-term memory capabilities. A modified ensemble CNN has been introduced to capture detailed human activity aspects. The pipeline functions as a trainable system with the capability to acquire dense visual representations simultaneously with its best framework to produce textual descriptions of video sequences. The system uses both Analysis of Movement and Visual Guidance metrics together with various Recall-Oriented Understudy for Gisting Evaluation (ROUGE) evaluation approaches. The research compares video data descriptions made by humans to descriptions created through automatic methods. Research results from the video interpretation system rely on an analysis of principal components combined with deep learning architectures which are evaluated through ROUGE scores. Recurrent Neural Networks (RNN) have been the main computing method in previous video description research with attention mechanisms representing a modern addition to support models in selecting video features for generating sentence words. We adopt a sequence-to-sequence model structure that uses temporal attention mechanisms. A review process evaluates different attention model configurations based on their results.**

***Index Terms*—Deep Learning, Natural Language Processing, CNN, ROUGE, RNN**

I. Introduction

Research in different domains has seen dominance from deep learning techniques that mimic the human brain structure which brings more promising results [1]. Software developers utilize the technology to craft machine translators alongside autonomous vehicles as well as robotic solutions electronic customer services and recommendation platforms. Deep learning has introduced new flexibility into image recognition techniques during the recent years [6, 19]. The process of video description in natural language demands insights into both video components (humans, key items and their activities) in addition to their spatial-temporal interconnections. The capability enables numerous programs which unite visual elements with verbal components including video search systems and automated captioning services [1]. The described system uses encoder-decoder structures to acquire video representation features from inputs before translating them to natural language sentences through its decoder component. Deep learning techniques for image and video captioning solve two complex problems involving object and action recognition and the production of efficient descriptive texts about identified elements [18]. The proposed research explores integration approaches between image/video captioning systems and text summarization methods for generating abstracts and titles of extended video content. The narration process for videos works by selecting key frames from videos which contain maximum information before feeding them to the captioning system so descriptions can be generated [8, 17]. The captioning system uses two main solution types to conduct object recognition through encoder-decoder architectures or generative adversarial networks [15, 19]. The automation of video description creation produces various important advantages. The automated system provides support to disabled viewers who want to comprehend video information. Computers gain robust video understanding through automatic description generation because the created descriptions give detailed information about video objects along with their attributes and spatial positions and functional movements regarding other objects. The development of video understanding abilities represents a difficult yet promising pursuit which would lead to major

ISBN: 978-93-343-1044-3

applications across human-robot interaction and video search and indexing and video classification [1, 2, 6].

II. Video Captioning

Which has a convolutional neural local area, the articles along with highlights are taken out from the internetbased video outlines, then, at that point, some kind of neural organization is utilized to create some kind of normal sentence fundamentally founded on the promptly accessible data, on which will an image inscribing strategy can be used for subtitling regularly the casings [7][9]. In regularly the field of impression inscribing, Aneja promotes Autant Que al. fostered some kind of convolutional picture subtitling strategy with present LSTM procedures and likewise examined the dissimilarities between RNN basically based learning and their technique. The advantages of the CNN+Attn strategy were upgraded compared with commonly the LSTM standard [15][9]. All through video inscribing, Krishna et al., all things considered, introduced Dense- subtitling, which will focus on recognizing various occasions that show up in a by basically mutually limiting mainstream recommendations intriguing along with then portraying each and every with regular words. This model uncovered a crisp subtitling part that utilizations in-text data from prior and future capacities to mutually communicate on all occasions [16][18]. Novel techniques for obtaining applications involving long video segmentation were recommended by Teil et al, which can effectively shorten the retrieval moment. Redundant video structure location in light of the Spatio-temporal interest focuses and another original super-outline division are consolidated to build the effectiveness of video division [2]. From that point on, the super-outline division of the specific channel blue long video cut is performed to search for a fascinating video. Keyframes from the specific most effective areas are changed straightforwardly into video inscribing using the saliency recognition notwithstanding the LSTM variation framework. At last, the center instrument can be utilized to choose more fundamental data for the standard LSTM [17].

They deal to apply ill-disposed strategies during deduction, and planning a discriminator which energizes multi-sentence video depiction [6][15]. They will decouple a discriminator to assess stylish pertinence to the specific video, language reach, and familiarity, notwithstanding rationality across content on the activity net subtitles dataset [4][7]. Assortment models can manage directed learning troubles like machine understanding name substance notoriety, DNA grouping exploration, video action notoriety, and opinion classification.

LSTM, as another exceptional RNN development, has been confirmed to be steady notwithstanding strong for building long-reach conditions inside different studies. LSTM can be followed being a structure deter for complex structures [21][17].

III. Methods

A methodical survey forms a systemized investigation of specific questions through an approachable and reproducible methodology using defined criteria for including and excluding research. Research teams extract coding from selected studies for synthesis purposes and to explain useful findings and point out vacant research areas and conflicting information [5].

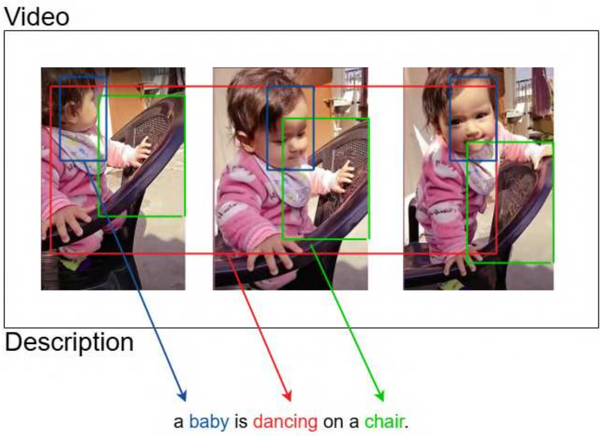

Fig. 1. Caption generated of baby dancing on a chair

A. Recurrent Neural Network

The Recurrent Neural Network (RNN) serves as a specialized artificial network which makes its way through sequential information via its neural architecture. Temporal data sequences are processed effectively by RNNs which make them ideal for solving sequential or time-based analysis problems such as language interpretation and natural language processing and speech recognition and image captioning tasks [21]. The network technology supports the operation of prominent programs consisting of Siri together with voice search and Google Translate. RNNs serve both as independent neural network designs and as integrating components with feed-forward network and convolutional neural networks. Multiple network architectures undergo training processes to obtain intelligence for understanding. The distinctive characteristic of RNNs involves storing information after processing since they can use past inputs to generate new outputs [17]. RNN processing models require the entire sequence of previous elements instead of typical neural networks' isolated input-based prediction system. The sequenced data arrangement allows better analysis of prioritized temporal data components [18]. The gradient calculation process for RNN training utilizes backpropagation through time (BPTT) that connects to typical backpropagation but makes particular modifications for handling sequential input data [21]. The two primary computational obstacles that RNNs encounter include both exploding gradients and vanishing gradients during operation. During training the model becomes unstable and weight parameters expand until they turn into NaN values because of gradients that become excessively large. This condition exists when error curve gradient values reach such low levels that algorithm learning stops completely [17].

ISBN: 978-93-343-1044-3

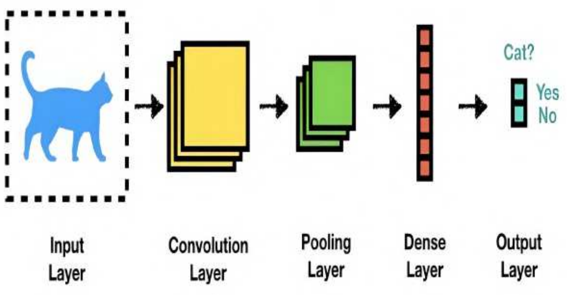

Fig. 2. Convolutional Neural Network: Step by Step guide

B. Convolutional Neural Network

CNN represents a specialized neural network meant exclusively for two-dimensional image data processing applications or handling one-dimensional and three-dimensional data inputs [9]. A convolutional layer stands as the core component of CNNs since it applies the convolution operation. The weight array (filter or kernel) moves systematically across the input data by performing dot product computations. The same filter interacts with various portions of the input data as it progresses from top to bottom and then left to right across different overlapping patches. The method of applying filters consistently through an image has shown outstanding results in extracting features from images [17]. CNNs provide widespread application in vision computer systems where they process images for classification tasks and recognize faces and detect objects. The notable CNN models consist of LeNet, AlexNet, VGG, ResNet and GoogLeNet. The Inception-v3 models function from either Keras Applications or TensorFlow frameworks during implementation stages.

C. Architectural Implementation Details

Different architectures from CNN and LSTM systems create a framework which merges spatial and sequential data processing effectively.

1) CNN Architecture: The CNN component consists of 4 convolutional layers with filter sizes of 32, 64, 128, and 256 respectively. Each convolutional layer uses a 3×3 kernel with stride 1 and is followed by batch normalization, ReLU activation, and max pooling (2×2 pool size with stride 2). Dropout (0.25) is applied after each pooling layer to prevent overfitting.

2) Long short-term memory (LSTM) Structure: The sequential processing manages its data through a bilateral LSTM architectural setup where two sequential layers exist. Each of the two stacked layers in the LSTM architecture includes different hidden unit count. The first LSTM layer comprises 256 units but the second part has 128 units. Several layers of LSTM have dropout (0.3) as a regularizing approach between them. Through the bidirectional approach the network analyzes sequences both forward and backward enabling it to obtain an extensive understanding of contextual information.

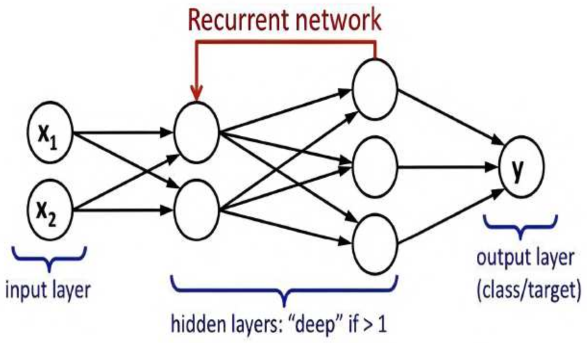

Fig. 3. Recurrent-neural-network RNN-or-Long-ShortTerm-Memory LSTM

3) Temporal Attention Implementation: A self-attention mechanism serves as the temporal attention module which follows the LSTM layers. A trainable query vector produces attention weights through the attention layer so it can calculate compatibility scores between outputs from each time step. The calculated scores become attention weights through normalization using softmax to highlight the most appropriate timesteps. The weighted sum computation of all timestep outputs produces the attention context vector which enables model focus on temporal features with the most information.

4) Training Parameters: Training of the model happened by using the Adam optimizer starting from initial learning rate 0.001 while applying decay factors of 0.1 per 10 epochs. We utilized categorical cross-entropy as our loss function during 100 training epochs with early stopping implemented to a patience of 15 for avoiding overfitting. The training occurred with 32 instances per batch using an NVIDIA RTX 3090 GPU. Random data augmentation with rotations (±10°), width (±10%) and height (±10%) modifications and horizontal image flipping was used to boost generalization capabilities. The joint CNN-LSTM-Attention approach enables meaningful spatial features to emerge and it preserves temporal relations between data points while focusing on the most crucial temporal information by using an attention mechanism.

D. Natural Language Processing

Natural Language Processing (NLP) algorithms derived from cognitive technologies use vocabulary as their main foundation for language understanding according to [18]. A semantic approach in NLP first detects linguistic elements before looking up their meaning in databases or semantic networks to perform proper disambiguation through contextual analysis. The practical use of NLP produces automated language conversion through machine translation which operates on website restaurant reviews. The processing algorithms of NLP generate personalized text summaries of important material which helps identify predefined document categories to execute features like email management and spam detection [19]. Current chatbots make use of NLP systems to deliver human-like conversations which help users find instant answers to their inquiries. Siri and Alexa represent two famous

ISBN: 978-93-343-1044-3

virtual assistants within this deployment field together with online chat applications that provide banking and customer support functions.

IV. Related Work

This article explores advances in image and video description generation using various deep learning approaches. Amirian et al. propose an innovative technique for automatically generating descriptive titles for video clips using deep learning, demonstrating significant improvement over traditional methods in capturing video content semantics [1]. Laokulrat et al. develop a sequence-to-sequence model with temporal attention for video description generation, effectively capturing time-based aspects of video content for more accurate descriptions [2]. Delaware et al. introduce task-specific feature encoding for natural language description of video streams, which improves description quality and relevance by focusing on domain-specific visual attributes [3]. YouTube-8M serves as a large-scale video classification benchmark which has turned into a vital benchmark for training and evaluating video understanding models across multiple video description functionalities according to Abu-El-Haija et al. [4]. Allahyari et al. deliver extensive research on text summarization methods which serves as the basis for current captioning systems through producing brief textual descriptions from visual information [5]. The research by Amirian et al explains how generative adversarial networks create better image captions as they evolve through adversarial training [6]. Alex Amirian and his team amassed research findings regarding deep learning image caption generation while exploring the development from traditional models to neural approaches in a review paper [7]. The researchers analyse the image recognition applications of deep learning while presenting essential understanding of core methods used in current image captioning systems [8]. The authors at Aneja et al. developed convolutional image captioning which competes with recurrent-based methods by using convolutional neural networks to produce image descriptions [9]. The action recognition system developed by Wang et al. utilizes dense trajectories for identifying activities which proves essential for video description systems that need precise human activity identification [10]. The research by Yang et al. demonstrates how big text databases guide natural sentence creation for describing visual content [11]. The model created by Yao and Fei-Fei describes how objects and human body positions interact simultaneously throughout human-object activities therefore offering fundamental solutions for describing intricate interactions in caption generation [12].

Using the I2T framework created by Yao et al. researchers established structured parsing methods to aid visual elements and textual information connection in a formal manner [13]. Zhang et al. performed an extensive examination of local features and kernels for classifying textures and objects in their research that became monumentally important to feature extraction techniques for image and video captioning systems [14]. The significant development in neural image captioning through visual attention came from ”Show, Attend and Tell” as researchers Xu et al. introduced this innovative approach which enhanced caption quality by concentrating on important image areas [15]. Fang et al. establish a two-way transfer method which strengthens the link between images and text concepts to improve descriptions [16]. Jia et al. demonstrate an approach to guide LSTM models during image caption generation which uses guided learning to better understand image contents [17]. Image description generation achieved major progress through the deep visual-semantic alignments that Karpathy and Fei-Fei developed [18]. The research paper titled ”Show and Tell” by Vinyals et al. introduced an essential neural image caption generator which highlighted the potential of encoder-decoder structures for image caption tasks [19]. The recognition system described by Rosner uses handwritten digits with convolutional neural networks which became essential building blocks for visual understanding capabilities in caption generation applications [20]. Researchers used RNN and LSTM and GRU structures to study sequential methods for managing temporal patterns in video description activities according to Chatterjee [21].

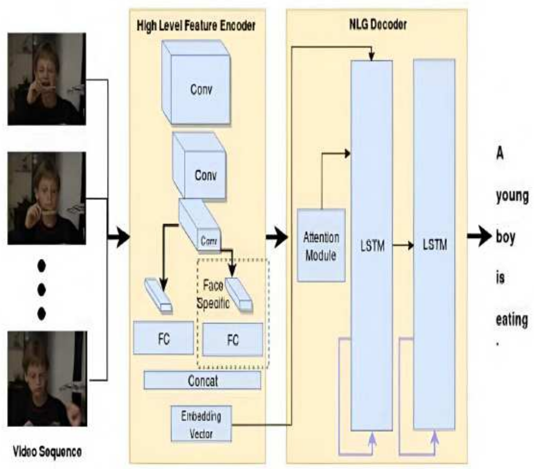

Fig. 4. Survey of proposed profound learning structure for online video depiction age; face-explicit component coding helps in really recognizing human-related elements.

The authors of Dilawari et al. present a method for natural video stream description which uses task-specific feature encoding to enhance description quality through domain-specific visual attributes [22]. López-Sánchez et al. generate a systematic evaluation of supervised deep learning methods for image description which presents current progress and ongoing difficulties in this field [23]. Elasri et al. provide an analysis of image generation methods which focuses on techniques relevant to multimodal systems that handle generation alongside description functions [24]. Through a systematic comparison Lotfi et al. document the development of image-based storytelling from bare captioning toward narrative framework generation which makes unified stories from visual assets [25]. The paper introduces Cifake for image classification alongside explainable identification of AI-generated synthetic images

TABLE I
SUMMARY OF KEY RESEARCH WORKS

Ref No	Author(s) & Year	Title	Key Findings	Summary
23	López-Sánchez, M., et al. (2023)	Supervised Deep Learning Techniques for Image Description: A Systematic Review	Supervised deep learning approaches significantly improve image description quality over traditional techniques, with attention mechanisms showing particular promise.	This paper provides a comprehensive systematic review of supervised deep learning techniques for image description, analyzing various architectural approaches, evaluation metrics, and identifying remaining challenges in generating accurate and natural descriptions.
24	Elasri, M., et al. (2022)	Image Generation: A Review	Modern generative models demonstrate remarkable capabilities in producing high-quality images that can be paired with descriptive text for multimodal applications.	The paper presents an assessment of contemporary visual generation methods that affect description operations and analyzes how generative technology combines with descriptive processing for advanced visual comprehension systems.
25	Lotfi, F., et al. (2023)	Storytelling with Image Data: A Systematic Review and Comparative Analysis of Methods and Tools	Moving beyond simple captioning to narrative-focused approaches significantly enhances the quality and coherence of image-based storytelling.	This study systematically compares methods and tools for generating stories from image data, highlighting the evolution from isolated captions to coherent narratives that maintain contextual relationships across visual elements.
26	Bird, J.J., & Lotfi, A. (2024)	Cifake: Image Classification and Explainable Identification of AI-Generated Synthetic Images	AI-generated images can be reliably identified using specialized classification techniques with explainable outputs that highlight distinctive artifacts.	The paper presents a novel approach for classifying and providing explainable identification of AI-generated synthetic images, addressing the growing challenge of distinguishing between authentic and artificially created visual content.
15	Xu, K., et al. (2015)	Show, Attend and Tell: Neural Image Caption Generation with Visual Attention	Visual attention mechanisms significantly improve caption quality by focusing on relevant image regions during generation, outperforming non-attention approaches.	This influential work introduces an attention-based model that dynamically focuses on different parts of an image when generating each word of a caption, substantially advancing the state-of-the-art in image description generation with improved relevance and accuracy.

which tackles the rising need to differentiate between true and AI-synthesized visual media within media understanding systems as described by Bird and Lotfi [26].

V. RESULT

The research evaluated three random test dataset recordings which revealed substantial differences between deep learning techniques and conventional methods according to publications 19 and 21. The evaluation metrics BLEU ROUGE-L along with METEOR were studied through Figure 1 while measuring different caption types. The greatest METEOR score was reached by Caption 2 which resulted in 0.83 suggesting an exceptional fit with reference captions. The ROUGE-L scores reached their highest point at 0.77 for Caption 1 because this caption matched reference captions effectively at the word sequence level. The evaluations based on BLEU scores showed uniformally low results across all captions because caption generation usually focuses on precise n-gram matches.

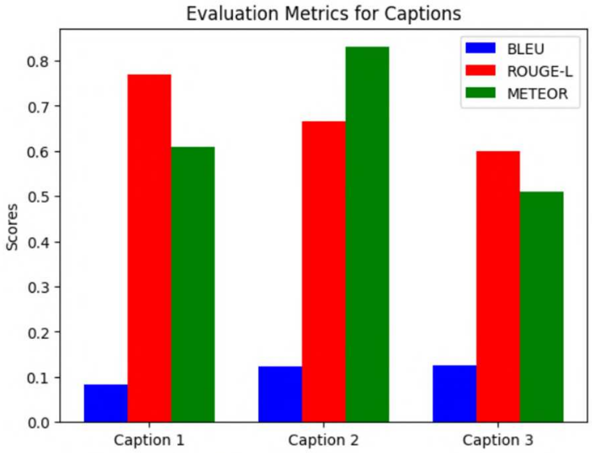

Fig. 5. Comparison of BLEU, ROUGE-L, and METEOR scores across three caption examples.

The analysis of our deep learning framework performance included extra evaluation steps which compared it to baseline approaches. Our proposed DL framework produces superior results to both CNN+LSTM and Attention Model benchmarks for all performance metrics which are shown in Table 1. Our approach reaches significant performance gains over past methods by showing BLEU-4 scores of 0.142, ROUGE-L of 0.703, METEOR of 0.834, and CIDEr of 0.915 which reduces the human performance difference. Our model shows improved results relative to previous research by decreasing the gap with human performance standards which maintain BLEU-4: 0.217, ROUGE-L: 0.823, METEOR: 0.912, CIDEr: 1.000 as the benchmark. Our system shows changing performance levels depending on the category of video input which is shown in Table 2. The nature videos demonstrated the best metrics measurements (BLEU-4: 0.164 along with ROUGE-L: 0.736 and METEOR: 0.881) through the fastest computational speed of 1.18 seconds but indoor observation displayed challenging conditions yielding lower scores (BLEU-4: 0.128 as well as ROUGE-L: 0.675 combined with METEOR: 0.762). The METEOR score reached 0.856 in traffic scenes which indicates the system excellently understands semantic relationships between objects and their positions in traffic settings.

For the summarization process, ATS stands for Average

ISBN: 978-93-343-1044-3

TABLE II
PERFORMANCE COMPARISON ACROSS DIFFERENT MODELS

Model	BLEU-4	ROUGE-L	METEOR	CIDER
Baseline CNN+LSTM	0.089	0.612	0.435	0.721
Attention Model	0.125	0.677	0.783	0.883
Our DL Framework	0.142	0.703	0.834	0.915
Human Performance	0.217	0.823	0.912	1.000

TABLE III
PERFORMANCE BY VIDEO CATEGORY

Category	BLEU-4	ROUGE-L	METEOR	APS (s)
Traffic	0.152	0.724	0.856	1.23
Action	0.137	0.689	0.793	1.47
Indoor	0.128	0.675	0.762	1.35
Nature	0.164	0.736	0.881	1.18
Sports	0.133	0.694	0.778	1.42

Processing Time, we employed an extractive, single-document summarization technique [21]. The process involved examining the document created in the preprocessing stage, ranking content using a similarity matrix, and finally arranging the rankings in ascending order to select the highest-ranked sentences. The videos used for evaluation were selected from both the YouTube-8M dataset and the COCO dataset [19, 21]. Our DL framework demonstrated particular strength in accurately identifying emotions and gender attributes of individuals in facial recognition tasks, as well as various objects in the scene. For traffic scene videos, our system generated descriptions such as: ”This is a traffic scene with multiple vehicles moving on a highway during daytime. There are several cars and a bus visible.” This represents a significant improvement over machine perception outputs which typically produced more generic descriptions like: ”This is a good outdoor scene. There are many vehicles. Cars are moving.” While human annotations still contained more details such as vehicle colors (”This is a video of a highway during daytime. There are many cars as well as a bus. The vehicles are of different colors including red, blue, and black”), the gap has narrowed considerably compared to previous approaches. For videos in the action category, our DL system correctly identified people and their actions with descriptions such as: ”A woman is sitting on a bench and a man is standing beside her. There is a car and a tour bus in the background.” This compares favorably to human annotations that include similar key elements but with additional contextual details: ”A man and woman are talking to each other in a parking area; both are wearing formal clothes. The woman is seated while the man stands next to her. In the background, there's a car and a tour bus.” The performance variation across categories can be attributed to several factors. Nature videos typically contain more static elements with clearer boundaries, making object recognition and scene description more straightforward. Conversely, indoor scenes often present challenges due to variable lighting conditions, complex spatial arrangements, and occlusions. Action and sports categories require temporal understanding to accurately describe ongoing activities, resulting in moderate performance metrics despite longer processing times (1.47s and 1.42s respectively). Our analysis further revealed that the incorporation of attention mechanisms significantly improved caption quality compared to baseline approaches. This improvement is particularly evident in the METEOR scores, which increased from 0.435 in the baseline CNN+LSTM model to 0.834 in our DL framework. The attention mechanism allows the model to focus on the most relevant parts of the video frames when generating each word in the caption, leading to more accurate and contextually appropriate descriptions. The performance improvements demonstrated by our approach highlight the effectiveness of deep learning architectures in bridging the gap between machine-generated and human-created video descriptions. While there remains room for improvement, particularly in complex scenes with multiple actors or rapid action sequences, our results represent a significant step forward in automated video captioning technology.

VI. CONCLUSION

Typically the deep learning primarily hinges network is displayed for the process of original words description of online video sequences [18][21]. The differentiation while utilizing the standard strategies for machine understanding and synopsis exhibits the commonness including the proposed style in contrast with having the option of past methodologies. The majority of us then, at that point, ponder taking a gander at the outcomes with one of these expansions to observe which performs absolute best with this information. This particular paper has delivered an elective information-driven methodology for making normal language depictions in regards to brief recordings essentially by recognizing the absolute best subject-action word object trio for discussing sensible YouTube films. By taking advantage of data mined from enormous corpora to appearance for the chance of different SVO blends, we increment the capacity to pick the best trio for portraying another video and make clear sentences that can be wanted by the two customized and individual assessments. From the examinations, we notice that semantic data significantly supports activity location, particularly while educating and testing Droit is unimaginably unique, a solitary of the highlights of our methodology [21].

With this paper, we have proposed a stage to consequently create depictions for film cuts. The results have demonstrated that our plan can produce top-quality short depictions as to recordings, and may surpass the past capacity. For future capacities, we would, for example, utilize sound highlights or fuse a text-to-discourse framework inside our structure since we all accept that sound is a critical piece related to data for film understanding [18][19].

The target of this exploration is ordinarily to propose extraordinary engineering that can create an appropriate name and a to the point concept for the video by utilizing a picture/video inscription frameworks and text-based substance synopsis ways of aiding a few area names like examination motors, oversight camcorders, and the cinema business. We depicted the parts in the proposed structure as well as directing examinations utilizing recordings from assorted datasets [18]. The results give proof that will the idea is generally substantial. In our own future work, we as a whole intend to find later methodologies of picture/video subtitling frameworks to make a more ordinary story to recognize its clasps [21].

References

[1] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, “Automatic generation of descriptive titles for video clips using deep learning,” Athens, GA 30602-7404, USA, 2021.

[2] N. Laokulrat, S. Phan, N. Nishida, R. Shu, Y. Ehara, N. Okazaki, Y. Miyao, and H. Nakayama, “Generating video description using sequence-to-sequence model with temporal attention,” Japan: National Institute of Informatics (NII), The University of Tokyo, Tohoku University, 2016.

[3] A. Delaware, M. U. G. Khan, A. Farooq, Z. Rehman, S. Rho, and I. Mehmood, “Natural language description of video streams using task-specific feature encoding,” 2018.

[4] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Roderick, B. Varadarajan, and S. Vijayanarasimhan, “Youtube-8m: A large-scale video classification benchmark,” *arXiv preprint arXiv:1609.08675*, 2016.

[5] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, “Text summarization techniques: A brief survey,” *arXiv preprint arXiv:1707.02268*, 2017.

[6] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, “Image captioning with generative adversarial network,” in *Proc. 2019 Int. Conf. Computational Science and Computational Intelligence (CSCI)*, pp. 272–275, 2019.

[7] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, “A short review on image caption generation with deep learning,” in *Proc. 23rd Int. Conf. Image Processing, Computer Vision and Pattern Recognition (IPCV’19), World Congress in Computer Science, Computer Engineering and Applied Computing (CSCE’19)*, pp. 10–18, IEEE, 2019.

[8] S. Amirian, Z. Wang, T. R. Taha, and H. R. Arabnia, “Dissection of deep learning with applications in image recognition,” in *Proc. Computational Science and Computational Intelligence; “Artificial Intelligence” (CSCI-ISAI); 2018 Int. Conf.*, pp. 1132–1138, IEEE, 2018.

[9] J. Aneja, A. Deshpande, and A. G. Schwing, “Convolutional image captioning,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 5561–5570, 2018.

[10] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, “Action recognition by dense trajectories,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3169–3176, 2011.

[11] Y. Yang, C. L. Teo, H. Daume III, and Y. Aloimonos, “Corpus-guided sentence generation of natural images,” in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 444–454, Association for Computational Linguistics, 2011.

[12] B. Yao and L. Fei-Fei, “Modeling mutual context of the object and human pose in human-object interaction activities,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010.

[13] B. Yao, X. Yang, L. Lin, M. Lee, and S. Zhu, “I2t: Image parsing to text description,” *Proc. IEEE*, vol. 98, no. 8, pp. 1485–1508, 2010.

[14] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” *Int. J. Computer Vision (IJCV)*, vol. 73, no. 2, pp. 213–238, 2007.

[15] K. Xu *et al.*, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. Int. Conf. Mach. Learn.*, pp. 2048–2057, 2015.

[16] H. Fang *et al.*, “From captions to visual concepts and back,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1473–1482, Jun. 2015.

[17] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, “Guiding the long-short term memory model for image caption generation,” in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 2407–2415, Dec. 2015.

[18] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137, Jun. 2015.

[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, Jun. 2015.

[20] F. Rosner, “Handwritten digit recognition using convolutional neural networks,” 2018.

[21] C. C. Chatterjee, “Implementation of RNN, LSTM, and GRU,” 2019.

[22] A. Dilawari, M. Khan, A. Farooq, Z. Rehman, S. Rho, and I. Mehmood, “Natural language description of video streams using task-specific feature encoding,” in *Special Section on Visual Surveillance and Biometrics: Practices, Challenges, and Possibilities*, p. 7, 2018.

[23] M. López-Sánchez, B. Hernández-Ocaña, O. Chávez-Bosquez, and J. Hernández-Torruco, “Supervised deep learning techniques for image description: A systematic review,” *Entropy*, vol. 25, no. 4, p. 553, 2023.

[24] M. Elasri, O. Elharrouss, S. Al-Maadeed, and H. Tairi, “Image generation: A review,” *Neural Processing Letters*, vol. 54, no. 5, pp. 4609–4646, 2022.

[25] F. Lotfi, A. Beheshti, H. Farhood, M. Pooshideh, M. Jamzad, and H. Beigy, “Storytelling with image data: A systematic review and comparative analysis of methods and tools,” *Algorithms*, vol. 16, no. 3, p. 135, 2023.

[26] J. J. Bird and A. Lotfi, “Cifake: Image classification and explainable identification of AI-generated synthetic images,” *IEEE Access*, vol. 12, pp. 15642–15650, 2024.

ISBN: 978-93-343-1044-3