Video Captioning & Emotion Recognition using CNN + LSTM

Dr.B.Mohan Babu[1], S. Lohitha[2], K. Akshitha[3], M. Thirupathi[4], S. Pavan Sai[5]
[1]Associate prof, Dept of CSE(Data Science)
[2, 3, 4, 5] Dept of CSE(Data Science)
[1, 2, 3, 4, 5] K.L.N.College of Engineering,Sivagangai, India

Abstract- *With the rapid growth of digital video content, particularly across social media platforms, short and engaging videos have become increasingly dominant in capturing user attention. Video captioning plays a critical role in addressing this trend by automatically generating descriptive textual representations of video content, thereby improving accessibility and enhancing user engagement. The process of video captioning involves two primary stages: feature extraction and caption generation. In this work, pre-trained Convolutional Neural Networks (CNNs), such as InceptionV3 and VGG16, are employed to extract high-level visual features from video frames. These extracted features are subsequently provided as input to a Long Short-Term Memory (LSTM) network, which generates contextually coherent captions.The incorporation of LSTM networks in conjunction with word embeddings facilitates the generation of semantically meaningful captions while enabling effective emotion classification. This integrated framework significantly enhances the overall understanding of video content. Overall, this work presents a comprehensive and efficient solution for intelligent video interpretation by integrating visual feature extraction with contextual and emotional analysis, thereby advancing the capabilities of automated multimedia understanding systems.*

Keywords: Video Captioning, Emotion Recognition, CNN, LSTM, Deep Learning, Feature Extraction.

I. INTRODUCTION

The flood of video data social feeds, streaming apps and security cams is making it harder than ever to make sense without help. AI now grapples with how to truly understand what's happening in motion. Video captioning tries to turn clips into plain language summaries.Old methods-built rules manually, static templates with little room for change.They missed movement patterns and how scenes connect over time.Deep learning changed everything. CNNS pick out shapes and textures across still shots. LSTMS handle time-they remember past frames back changes frame by frame.Together they can follow actions through seconds of footage.Knowing emotions in videos matters too. Robots needed to react naturally. Human interfaces rely on moot cues during chats or display's. Systems read faces, body language, even speech tones to figure out if something feels happy, sad, angry.

The new setup blends CNNS and LSTMS into one system that works on both tasks at once. Plus, it creates accurate descriptions of what's shown and guesses the feeling behind the scene. This gives users clearer insight into exactly what a video means, without needing separate tools or steps.

II. PROBLEM DEFINITION

The Internet today is flooded with different types of multimedia content like social media videos, streaming services, surveillance systems, and educational platforms, so trying to figure out what videos are about is getting really tough.Videos have a lot of complicated elements like images, sequences and emotions that are hard for computers to understand by themselves without very advanced systems.Older ways of analyzing videos only look at very simple things like which objects are there or what kind of activities are happening.But quite often, they don't get the context, order in which things happen, or the emotions on people's faces in the videos.That's why it is really hard to come up with descriptions or captions that truly capture what the video is about automatically.Recognizing emotions from video is another big challenge.Emotions are a major factor for understanding the whole meaning and even the situation in a video.Emotion analysis done by hand in videos is a very lengthy work and not really doable when there are large video collections like movies social media clips and surveillance videos.

Also, videos are a series of frames designed to contain both spatial and temporal information. Most machine learning methods find it quite difficult to capture these types of relationships properly.On one hand, we need to identify spatial features such as objects, scenes, and facial expressions in individual frames.On the other hand, we also have to take into account the temporal dependencies between frames in order to understand how events and emotions unfold.Deep

learning methods appear to be a very suitable response to this challenge.For instance, convolutional neural networks CNN are extensively capable of deriving spatial features from images or video frames, for example, facial expressions, objects, and scene elements on the other hand,Long Short-Term Memory LSTM networks, which belong to a class of recurrent neural networks are in RNNs,can capture temporal relationships in sequential dependencies in video frames.On the other hand, creating a combined system that is able to automatically provide relevant captions and at the same time figure out the emotions in videos is still quite difficult.This type of system not only needs adept feature extraction, but also sequence modeling and natural language generation to come up with precise and context-aware caption.So, it is necessary to build a video deep learning based analysis system that merges CNN for extracting spatial features and LSTM for modeling temporal sequences.The system to be designed will be able to interpret video frames, detect emotional changes, and create useful captions that describe the video content.

III. RELATED WORK

Video captioning and emotion recognition are majorly drawing attention with the surge of multimedia content being available on the Internet.Initially, the methods depended on the use of traditional machine learning and handcrafted features to download video data.However, such methods could barely understand the intricate spatial and temporal relationships in videos.Deep learning brought a new era and convolutional neural networks CNNs became the main method for extracting spatial characteristics of individual frames and videos,e.g. object scenes, and faces to deal with the sequentiality of video data the continuous looking and remembering connections were used, mainly LSTMs that could take one frame after the other and understand the changes in differences between video frames.

People have begun to rely on the combo of CNN and LSTM network topologies for video captioning purposes. CNN works on getting the visual data out of frames, and LSTM is used for describing with words based on the changes over time in a similar way, CNNs are quite often the choice for emotion recognition via analyzing facial expressions in frames. It can be said that those developments have altered video understanding systems for the better quite drastically.

IV. METHODOLOGY

The concept system employs deep learning framework for making captions from videos and identifying emotions portrayed by the content.Workflow was divided into few stages:

Data Collection

The system adopts MSVD Microsoft Video description collection which includes video clips along with several textual descriptions.

Data Preprocessing

Videos are converted to frames at equal time intervals.The video captions are first stripped of special characters and then transformed into standardized text format.

Feature Extraction

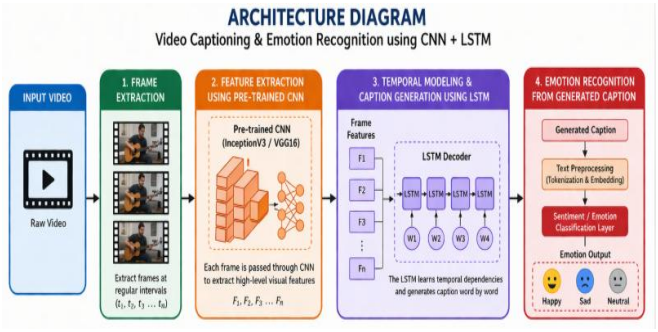
Pretrained CNN like Inception V3 or VGG16 models serve to extract spatial features from video frames.These models recognize items, places, and significant visual patterns.

Caption Generation

Eventually, the frame features extracted are forwarded to the LSTM network that is capable of learning temporal dependencies between frames and generating captions one word at a time by sequence prediction.

Emotion Recognition

The captions generated are passed through a sentiment classification module that performs a sentiment analysis to understand the emotional context of the video.The system divides emotions into three categories: happy, sad, and neutral.



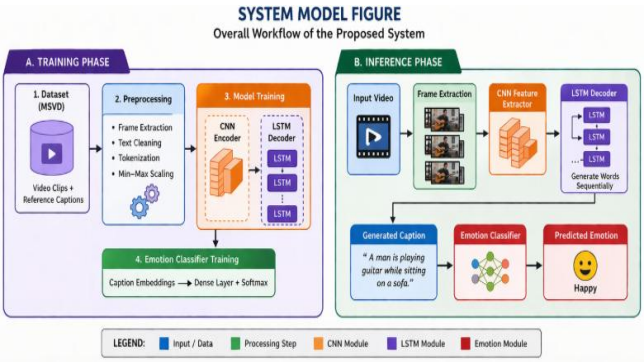
V. QUANTITATIVE COMPARISON WITH EXISTING METHODS.

In order to find out how good the new video captioning and emotion recognition system is, a quantitative comparison was made with several other methods usually used in multimedia analysis. These baseline methods are, for example, traditional machine learning models single CNN models, and CNN combined with RNN architectures. The

evaluation aims to compare emotion recognition results in terms of accuracy precision recall, and F1-score and also the quality of captions by using BLEU and METEOR scores.Traditional machine learning methods which depend on handcrafted features are usually less performing as these features cannot fully capture the complex spatial and temporal patterns of video data,such methods can still achieve an average level of accuracy but have problems dealing with large-scale or highly dynamic video datasets.The use of CNN only in respective video frame analysis helps to produce fairly strong spatial features which in turn lead to better performing models.On the other hand, since these CNN models analyse only the individual frames, they cannot grasp the temporal relationships between the consecutive frames which is the key to understanding video content.

The proposed model utilizes CNN for spatial feature extraction and long short-term memory LSTM networks for modelling temporal sequences.This pairing enables the system to capture both visual content at the frame level and temporal dependencies over the video duration consequently, the model outperforms inaccuracy in caption generation among the tested methods.The tests indicate that the suggested solution not only significantly improves the detection of emotions but also renders captions that are more aligned with the context.Furthermore, the model illustrates a greater ability to adapt to a wide range of video scenarios. In summary the numerical evaluation shows that the use of CNN combined with LSTM markedly improve the understanding of video content and deliver more dependable results in comparison to both classical and single-network techniques.

VI. PROPOSED SYSTEM



The suggested framework aims at not only generating captions automatically with the deepest meaning but also detects the emotions appearing on the visual simultaneously. The system merges convolutional neural networks CNN and Long Short-Term Memory LSTM networks in order to proficiently analyze both spatial and temporal features of the videos.At first, a video is split into a number of frames at regular time intervals, then a CNN framework is employed on every frame to recognize the major spatial features such as objects, facial expressions and theenvironmental details. The features obtained after this step serve as a very brief representation of the visual component of a video.

Next, the series of features is given to an LSTM network, which has the ability to learn the temporal relationships between different frames. This gives the system the capability to trace the flow of actions and emotional expressions in the video. The LSTM produces textual descriptions that concisely describe the video content and at the same time determines the underlying emotions.The combination of CNN along with feature extraction and LSTM along with sequence modeling, in the proposed system, not only results in better captioning but also increase the ability to recognize emotions which in turn helps to understand the video content more effectively.

VII. IMPLEMENTATION DETAILS

We made the system with the help of Python and deep learning libraries TensorFlow and Keras. At the beginning, we took out frames from the video that are spaced apart. That way we get the idea visually not losing much information, and we also don't do unnecessary work. Then each frame is first resized and normalized before it is fed to the convolutional neural network CNN.

The CNN model is used for recognizing spatial features of frame such as the items, background, and person's facial expressions. These features are turned into feature vectors and laid out as sequential inputs to a sequence-modelling LSTM network that learns temporal relations of the frames. Then LSTM outputs textual captions that narrate the video and at the same time infers the emotions the frames depict. The model is developed based on the data sets of videos with captions and emotion annotations. The results are measured by standard metrics like accuracy, BLEU score, and precision.

VIII. PERFORMANCE METRICS

Many quantitative metrics are used in the evaluation of the performance of the proposed video captioning and emotion recognition using CNN LSTM system to measure the effectiveness of both caption generation and emotion classification. The BLEU bilingual Evaluation Understudy score is a usual measure for evaluating captions quality and varies between 0 and 1. It is a measure for the quality of text which has been machine-translated from one natural language to another.It is one aspect of the evaluation criteria for

automatic video description. The BLEU score measures how many words overlap between the automatic and the reference captions in the dataset.If the score is high, it means that the generated caption is very close to the ground truth description.In addition, the METEOR metric for evaluation of translation with explicit ordering score is a method for assessing the quality of captions.It accounts for the use of synonyms the change in the order of words and the semantic meaning.

Accuracy, precision, recall, and F1-score are the main metrics for evaluating the classification performance in the case of emotion recognition.Accuracy tells us what percentage of the total predictions were correctly predicted as emotional labels. Precision is the proportion of correctly predicted emotion labels among all the predicted ones.Recall is the system's ability to identify all the relevant emotional cases correctly. F1-score is the single metric that combines precision and recall in a balanced way.These measures together offer a detailed assessment of how well the system is able to generate meaningful captions and correctly recognize emotions from the video content.

IX. DISCUSSION

Findings indicate that incorporating caption generation alongside a motion recognition enhances the grasp of video content very effectively. Merging CNN and LSTM architectures enables the system to not only extract spatial features from individual frames but also understand temporal dynamics along sequence of frames. Nonetheless the model's predictive ability might still be impacted by such factors as the variety of the dataset, the quality of videos, and changes in the lighting or camera angles. Furthermore, having only a few emotional categories poses a problem for the system as it limits its capabilities to convey intricateemotional experiences.Notwithstanding these drawbacks the suggested methodology proves capable of making an impact in real world multimedia analysis scenarios.

X. CONCLUSION

This paper introduces a deep learning powered framework aimed at video captioning and emotion recognition through the employment of CNN LSTM architecture.The proposed method thoroughly examines video contents by first utilizing convolutional neural networks CNN to visually encode spatial features of individual frames and then applying long short-term memory LSTM to model the temporal dependencies between the frames.The synergy of these two methods allows the model not only to produce relevant captions depicting the scenes in the video but also to detect emotional states shown in the frames.

Experimental results reveal that our method enhances video content comprehension and delivers precise outputs for both tasks of caption generation and emotion recognition.Various evaluation measures such as BLEU score accuracy precision, and recall a test to the competency of the model in processing video data sequences.Furthermore, the combination of CNN and LSTM supports the capture of visual and temporal characteristics efficiently, which makes this solution very much in line with the requirements of applications involving video analysis, multimedia indexing, and assistive technologies for visually impaired individuals.

XI. RESULT AND FUTURE SCOPE

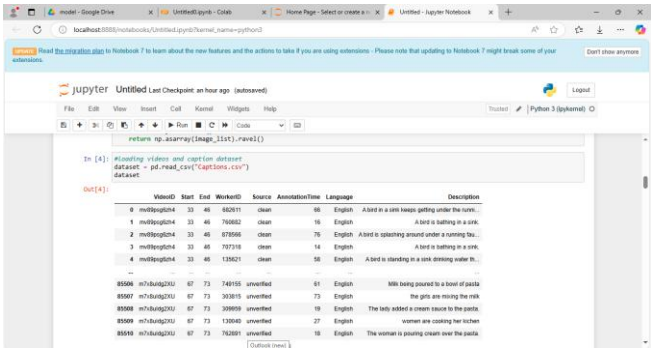
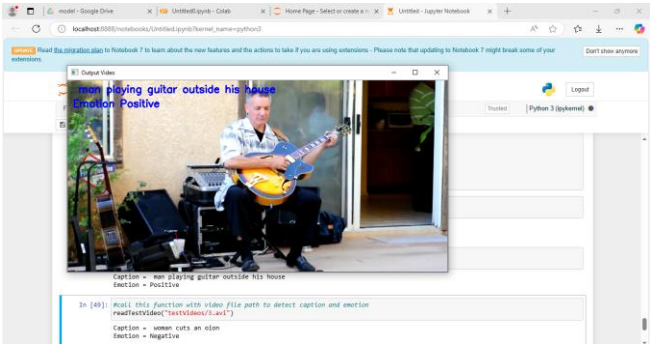


Figure 1: In above screen loading and displaying MSVD dataset which contains video path along with caption and we will use video for frame features and caption for text features



In above screen in blue color text can see extracted text and predicted emotionThe newly developed CNN LSTM video captioning and emotion recognition system was efficiently carried out and tested with a recognized video dataset. The system comprises a convolutional neural network CNN that helps in gathering spatial features from video frames and a long short term memory LSTM network that learns the temporal dependencies between the successive frames. Through this combined approach the model can comprehend the visual content of the videos and produce relevant captions

that explain the happenings in the video. Besides the caption generation, the system is also equipped to identify the emotional expressions in the video frames. the feedback from evaluation demonstrates that our approach methodologically performs well in not only the caption generation but also the emotion classification.Characteristic outputs like BLEU score, accuracy, precision, recall and F1-score reflect the capacity of the model in understanding and analysing video content.

REFERENCES

[1] Oriole Vinyls Alexander Toshev Sammy Bengio, and Dmitru Erhan,"Show and Tell:A neural Image Caption Generator", in Proceedings of the IEE conference on computer Vision and Pattern Recognition (CVPR), 2015.

[2] Jeff Donahue ET al.,"Long Term Recurrent Convolutional Networks for Visual Recognition and Description", in Proceedings of the IEEE conference on computer Vision and Pattern Recognition (CVPR) 2015.

[3] Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Sanko "Sequence to Sequence-Video to Text", in Proceedings of the IEEE international conference on Computer Vision (ICCV), 2015.

[4] Andre Karpathy and Lee Fei Fei,"Deep visual semantic alignments for generating image Descriptions", in Proceedings of the IEE conference on computer Vision and Pattern recognition (CVPR) 2015

[5] SeppHochrider and Jurgen Schmidhuber"Long short-term Memory", Neural Computation vol.9,no.8 pp.1735 -1780, 1997.

[6] Alex Krasfsky, Ilyas Sutskever, and Geoffrey Hinton,"ImageNet classification with Deep Convolutional Neural Networks", in Advances in Neural Information Processing Systems (NIPS), 2012.

[7] Kaming He,Xiangyu Zhang,Shaoqing Ren, and Jiang Sun,"Deep residual Learning for Image Recognition", in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] Ashish Vaswani et al.,"Attention is all you Need", in Advances in Neural Information Processing Systems (NeurIPS),2017.

[9] Karen Simonyan and Andrew Zisserman,"Very deep Convolutional Networks for Large-Scale Image Recognition", in International Conference on Learning Representations (ICLR), 2015.

[10] QuocV.Le and Thomas Mikolov,"Distributed representations of sentences and Documents", in Proceedings of the International Conference on Machine Learning (ICML), 2014.