Image Caption Generation using Deep Learning For Video Summarization Applications

Mohammed Inayathulla[1], Karthikeyan C[2]
Research Scholar, Department of Computer Science and Engineering[1]
Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India[1]
Associate Professor, Department of Computer Science and Engineering[2]
Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India[2]

***Abstract*—In the area of video summarization applications, automatic image caption synthesis using deep learning is a promising approach. This methodology utilizes the capabilities of neural networks to autonomously produce detailed textual descriptions for significant frames or instances in a video. Through the examination of visual elements, deep learning models possess the capability to discern and classify objects, scenarios, and actions, hence enabling the generation of coherent and useful captions. This paper presents a novel methodology for generating image captions in the context of video summarizing applications. DenseNet201 architecture is used to extract image features, enabling the effective extraction of comprehensive visual information from keyframes in the videos. In text processing, GloVe embedding, which is pre-trained word vectors that capture semantic associations between words, is employed to efficiently represent textual information. The utilization of these embeddings establishes a fundamental basis for comprehending the contextual variations and semantic significance of words contained within the captions. LSTM models are subsequently utilized to process the GloVe embeddings, facilitating the development of captions that keep coherence, context, and readability. The integration of GloVe embeddings with LSTM models in this study facilitates the effective fusion of visual and textual data, leading to the generation of captions that are both informative and contextually relevant for video summarization. The proposed model significantly enhances the performance by combining the strengths of convolutional neural networks for image analysis and recurrent neural networks for natural language generation. The experimental results demonstrate the effectiveness of the proposed approach in generating informative captions for video summarization, offering a valuable tool for content understanding, retrieval, and recommendation.**

Keywords*—Video summarization; deep learning; image caption synthesis; densenet201; GloVe embeddings; LSTM

I. INTRODUCTION

Making captions for images is an interesting and useful area of computer vision and natural language processing. The process involves developing algorithms and models that enables machines to provide descriptive and contextually appropriate textual captions for images [24] [25]. This technological advancement serves to connect visual and textual data, so enabling deeper understanding of image content and creating opportunities for diverse applications [1]. The field of image captioning is gaining considerable interest owing to its capacity to boost image accessibility, assist individuals with visual impairments, automate content creation, and enhance image retrieval systems [2]. Especially in video summarization, image caption generation is a potent tool with uses that go beyond individual images. The goal of video summarization is to reduce long videos' main points to more manageable chunks so that viewers may quickly understand the main points without having to watch the full thing. Image caption generation is essential in this situation [3].

The process of video summarization often entails splitting a video into a series of frames, which are effectively separate images. After that, approaches for image captioning are used to these frames, which results in the generation of written descriptions for each frame. There are many ways that image caption creation might be used to the process of video summarization [4]. It may be used in news collection, which provides consumers with the ability to quickly interpret the most important aspects of news broadcasts or events. In educational settings, it can facilitate efficient learning by providing concise summaries of lengthy video lectures. It may be used by content makers to generate appealing video teasers or trailers, and it also has potential applications in surveillance and security, where it might assist analysts in more quickly reviewing video material [5]. The use of image captioning in video summarization not only makes the process of processing material more streamlined, but it also improves searchability and the ability to retrieve certain video portions. This convergence of computer vision and natural language processing puts us one step closer to developing video summarization tools that are more efficient and user-friendly. Due to these challenges, the field of video summarization [6] has seen the emergence of deep learning as a very promising approach. Deep learning, specifically deep neural networks, has an impressive capacity to autonomously acquire complex features and patterns from unprocessed data [27][28]. Furthermore, it can efficiently capture the temporal relationships present in video streams. The process involves the use of sophisticated neural networks and algorithms to examine video streams and identify frames that most effectively depict the information or occurrences inside the film. The frames that have been chosen are often known as "image captures" and function as succinct summaries of the video material [7]. This methodology exhibits a range of realistic implementations, including the retrieval of video material, security and surveillance operations, video search functionalities, and content analysis activities. The primary objective is to discern significant information promptly and effectively inside vast collections of video data. The use of deep learning techniques

enhances the efficacy and adaptability of the summarization process, enabling its application across diverse video kinds and areas [8]. It has an intrinsic capability to modify and generalize over a wide range of video genres, making it an attractive option for tackling the complex challenges presented by video summarization.

The objective of this research is to explore the integration of deep learning methodologies with video summarization, specifically emphasizing the production of picture snapshots. It investigates the capabilities of deep learning models in autonomously selecting and constructing cohesive image-based summaries of movies. It presents a valuable tool with wide-ranging applications in many fields such as surveillance, content analysis, and video search. The subsequent parts provide an overview of the most recent deep learning approaches; proceed throughout their benefits for video summarization, and show case studies and experimental findings that demonstrate how well these methods work to solve the problems associated with creating image captures from video streams. The organization of the paper as follows: Section II describes the literature survey and the proposed model is explained in Section III. The simulation results are discussed in Section IV. Finally, Section V concludes the paper.

II. Related Work

Hafiz Burhan Ul Haq et al. [9] proposed a deep learning based system for tailored video summarization. The suggested framework facilitates video summarization based on the Object of Interest (OoI), such as individuals, aircraft, mobile devices, bicycles, and automobiles. Sridevi et al. [10] demonstrated with the use of a deep convolutional neural network in each stream, to create a video summary by extracting the temporal and spatial information from a video. The use of a Two-dimensional Convolutional Neural Network (2D CNN) allows for the exploitation of spatial information, while a Three-dimensional Convolutional Neural Network (3D CNN) is employed to exploit temporal information in order to provide highlight scores for video segments. The fusion of segment ratings from each stream is used to identify highlight portions within the video. Moreover, the resulting highlight representation alone indicates the user's relative degree of interest in a movie. To train the deep convolutional neural network (DCNN) in each stream, a paired deep ranking model is used. The objective is to enhance the highlight score of the highlight segment relative to the non-highlight section via the optimization of the model. The segments that have been acquired are then used in the process of summarizing a video.

Obada Issa et al. [11] illustrated novel methodologies for addressing the issue of key frame extraction in the context of video summarization. The methodology used in the study involves the extraction of feature variables from the bit streams of coded films, which is then followed by an optional stepwise regression process aimed at reducing dimensionality. After extracting the features and reducing their dimensionality, novel frame-level temporal subsampling approaches are used, followed by training and testing using deep learning architectures. The frame-level temporal subsampling approaches rely on the use of cosine similarity and the application of PCA projections on feature vectors. Three distinct learning architectures are constructed by using LSTM networks, 1D-CNN networks, and random forests.

Xu Wang et al. [12] presented a novel deep summarizing network that incorporates auxiliary summarization losses in order to effectively tackle the aforementioned issue. The incorporation of an unsupervised auxiliary summarization loss module using LSTM and a swish activation function is proposed. This module aims to effectively capture long-term dependencies for video summarizing tasks. Furthermore, the proposed module can be seamlessly incorporated into diverse network architectures. The presented model is a novel unsupervised framework for deep reinforcement learning that operates independently of any explicit labels or user interactions. In addition, the suggested model has a low computational burden and may effectively be implemented on mobile devices, hence improving the mobile user experience and alleviating strain on server operations.

Rhevanth et al. [13] presented an effective video summarizing method that extracts essential frames from raw video input and analyzes visual and audio material. Mel-frequency cepstral coefficient (MFCC) extracts information from audio sources, whereas structural similarity index compares frames. Using the preceding two functions removes superfluous video frames. A deep convolution neural network (CNN) model refines the key frames to get a list of potential key frames that summarize the data. Gulraiz Khan et al. [14] suggested a method facilitating users in generating video summaries by using human and object attributes. Cryptographic hashes play a crucial role in the context of blockchain technology. These hashes are derived from condensed video blocks, serving as a means of summarizing the content. Subsequently, these hashes are signed and transferred over the blockchain network. The Cumulus blockchain method is used to safeguard the integrity of the video. The system facilitates distant users in obtaining tamper-proof, condensed video footage of their company locations or other critical properties, which can be accessed on their cellphones. Xiaoning Chen et al. [15] presented a method for video summarization (VS) that leverages the complementary nature of shallow and deep features. The suggested approach involves Multiview feature co-factorization based dictionary selection, which aims to use the shared information from both shallow and deep view features in VS. In order to effectively use the whole visual information of video frames, two view features are employed. Subsequently, the shared information between these two distinct views is extracted using coupled matrix factorization. This retrieved information is then utilized for the purpose of dictionary selection in the context of visual surveillance. Ke Zheng et al. [16] presented a video summarization generation model called DME-VSNet, which utilizes a multi-feature approach to extract various information from the video frames. This study incorporates three key variables: significance score, picture memory strength, and image entropy. In response to the issue of imprecise video shot segmentation, this study presents a video shot segmentation algorithm that utilizes the TransNet network. The system effectively partitions the original video into many shorter shots by identifying shot borders. The suggested model incorporates

three specific variables as inputs for this purpose. The video frame score is acquired inside the Multi-Layer Perceptron (MLP) architecture, and subsequently, the key frame is determined based on this score to provide a concise summary of the movie.

Balamurugan et al. [17] illustrated a model for anomalous event detection that combines a hybrid convolution neural network (CNN) with bi-directional long short term memory (Bi-LSTM). The model is designed to have decreased complexity. The proposed model incorporates a convolutional neural network that utilizes a pre-trained model to extract spatio-temporal features from individual frames within a series. These features are subsequently fed into a multi-layer bi-directional long short-term memory network, which is capable of accurately classifying abnormal events in complex surveillance scenes on the road. The fine-grained technique incorporates a hierarchical temporal attention-based LSTM encoder-decoder model to provide an enhanced video summarizing approach that effectively maintains critical information while optimizing storage capacity.

Sah, Ramesh Kumar et al. [18] proposed a framework that utilizes spatial and temporal aspects, including self-attention mechanisms, to select representative content from video sequences. The framework generates temporal proposals and employs supervised learning techniques using manually provided data from individuals or users. Current supervised approaches do not effectively address the temporal interest and its consistency. In addition, achieving temporal consistency requires the ability to anticipate the temporal suggestions of the video segment. The present study approaches the task as temporal action detection, whereby it aims to concurrently forecast the relevance score and placement of the segments. This is achieved by using an anchor-based system that creates anchors of different lengths to effectively identify intriguing ideas.

III. Proposed Methodology

A deep learning-based image caption generator is a framework that automatically generates informative text captions for images. By merging computer vision and natural language processing methodologies, this system is capable of comprehending and articulating the semantic content of an image in a manner that is comprehensible to humans.

A. Image Caption Generator Framework

The image caption generator framework consists of data collection, data presentation, model presentation, and training and validation phases. During the data preprocessing stage, images undergo several operations such as scaling, normalization, and augmentation to ensure uniform dimensions and improved feature representation. Captions/Textual descriptions undergo the process of tokenization, wherein they are segmented into individual units, and subsequently transformed into numerical representations. This conversion is commonly achieved through the utilization of word embedding techniques. The model architecture is implemented using a pre-trained convolutional neural network (CNN), an encoder, which is responsible for processing the image and extracting visual features at a higher level. On the other hand, a decoder, commonly implemented as a recurrent neural network (RNN), utilizes these extracted features to generate textual descriptions. During the training process, the model aims to reduce the disparity between the captions generated by the model and the actual captions by utilizing a loss function, such as cross-entropy. Fig. 1 shows the Image Caption Generator Framework.

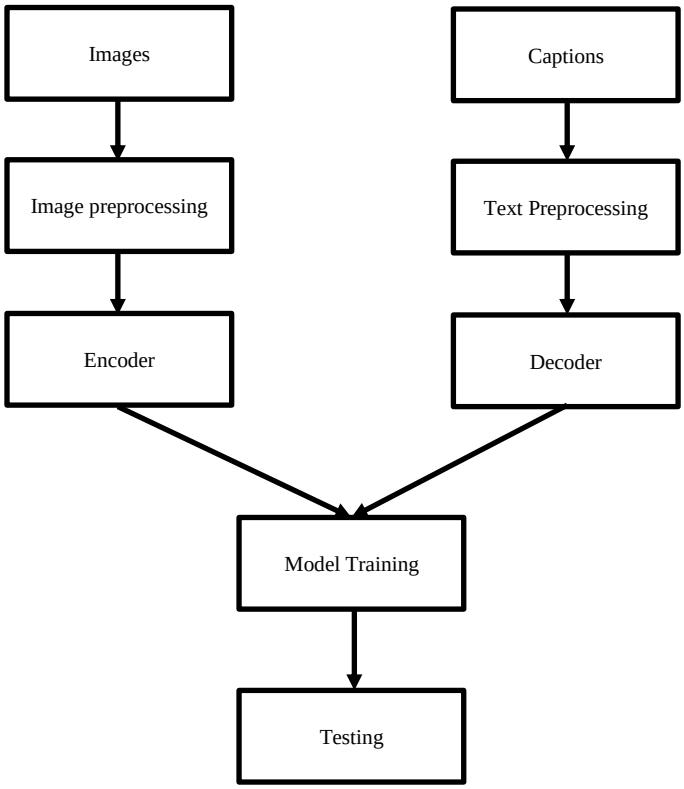

Fig. 1. Image caption generator framework.

1) Image preprocessing: Resizing: To maintain consistency, images in the dataset are frequently scaled to a fixed dimension (such as 224x224). The necessity of performing the resizing phase arises from the fact that deep learning models, particularly convolutional neural networks (CNNs), commonly necessitate input images to possess equal dimensions.

Normalization refers to the process of scaling image pixel values to a predetermined range, commonly denoted as [0, 1] or [-1, 1]. Normalization is a crucial step in data preprocessing that aims to establish a uniform range and mean for the input data. This process plays a significant role in enhancing the training process and facilitating the convergence of the model.

Data augmentation strategies are employed in order to enhance the variety of training data and bolster the resilience of the model. This may encompass various image processing processes such as rotation, cropping, flipping, brightness tweaks, and zooming. Data augmentation is a technique that aids in enhancing the generalization capabilities of a model towards images that have not been previously encountered.

2) Text preprocessing: Tokenization involves the process of breaking down captions, into individual words or sub words. Tokenization refers to the computational procedure of dividing a given text into distinct and meaningful parts, which might include individual words or sub word tokens. This process is commonly accomplished through the utilization of tools such as the Natural Language Toolkit (NLTK) or spaCy. Every token is representative of either a complete word or a fragment of a word.

Vocabulary creation involves extracting the tokens present in the dataset and organizing them into a comprehensive collection. This lexicon encompasses all distinct lexical units or morphological constituents found inside the captions. The size of the vocabulary is governed by the quantity of distinct tokens. Restricting the size of the vocabulary is crucial to maintain computational efficiency during training and inference, as excessively large vocabularies can lead to increased computational costs.

Padding and sequence length issues arise when dealing with captions, as they frequently vary in length. However, neural networks require input sequences of a set length. Consequently, it is possible to add a specific token (such as <PAD>) to sequences in order to achieve consistent length. The determination of a maximum sequence length is also utilized to appropriately truncate or pad sequences. Captions that are lower than the maximum allowable length are extended by adding padding, and captions that exceed the maximum length are shortened by truncation. Word embeddings are frequently employed to transform words into numerical representations. Pre-existing word embeddings, such as Word2Vec, GloVe, and FastText, can be employed to establish a mapping between words and compact vector representations. Embeddings can capture semantic links between words, hence enhancing the model's comprehension of the text. Special tokens, such as "<START>" to denote the initiation of a series and "<END>" to signify its conclusion, are incorporated into the tokenized captions. The utilization of tokens aids the model in acquiring knowledge regarding the appropriate instances to initiate and conclude the process of generating captions during the decoding phase.

Data preprocessing is a crucial step in preparing picture and text data for effective training of deep learning models. The preprocessed data is subsequently utilized to train the image encoder, which is typically a Convolutional Neural Network (CNN), and the text decoder, which is typically a Recurrent Neural Network (RNN) model, in the image captioning framework. The alignment between the data and the model architecture is crucial for effectively training the model and producing coherent image captions.

3) Image encoder: The function of the image encoder is to undertake the processing of the input image and extract significant features from it. Convolutional Neural Networks (CNNs) are frequently employed as image encoders.

a) Pre-trained CNN: A pre-trained (CNN) model, such as VGG, ResNet, or Inception, is employed as backbone for the image encoder. These models have already been trained to extract hierarchical and meaningful visual features from images using data from large-scale image classification tasks like as ImageNet.

b) Feature extraction: To extract features, the input image is processed through the pre-trained CNN. As the input data traverses the many layers of the CNN, distinct characteristics are identified and represented at varying degrees of abstraction. The properties serve to capture and represent relevant information related to the edges, textures, forms, and constituent components of the depicted image.

c) Dense layer: Dense layer or fully connected layer is employed to further transform the feature vector into a feature map that aligns with the desired input size for the text decoder. Embedding of Image Features: The image encoder produces a feature vector as its final output, which serves as a representation of the visual material contained inside the image. The feature vector serves as the initial hidden state for the text decoder. The process of embedding image features into a dense vector is commonly employed to ensure compatibility with the RNN decoder.

4) Text decoder: The text decoder reads the visual features and provides textual captions word by word. Text decoders in natural language processing (NLP) often employ Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks. The following is a comprehensive elucidation of the text decoder:

a) Initial state for the text decoder is derived from the feature vector obtained from the image encoder. This initializes the decoder by utilizing a reference point to generate captions that are derived from the visual characteristics of the image.

b) Embedding layer: The input word tokens or sub word tokens obtained from the preprocessed captions undergo a process of passing via an embedding layer. The process involves the mapping of each word to a dense vector representation, often of a predetermined size. The acquisition of these embeddings takes place during the training process.

c) Recurrent layers: The fundamental component of the text decoder consists of the recurrent layers, specifically the Long Short-Term Memory (LSTM) . The layers receive the embedded word representations as input and iteratively update their hidden states. During each iteration, the decoder generates a prediction for the subsequent word in the caption. The hidden state is modified by using information from both the previously created words and the image features.

d) Output layer: The output layer of the decoder generates a probability distribution across the vocabulary of words at each time step. The generation of this distribution is accomplished by applying a softmax layer on the hidden state. The subsequent word in the caption is selected based on its highest probability. The integration of an image encoder and text decoder inside a sequential framework constitutes the fundamental component of the image captioning model. The objective of training this model is to reduce the disparity between the captions generated by the model and the ground truth captions obtained from the dataset. The model acquires

the ability to produce coherent and contextually appropriate captions for a diverse array of images.

B. Proposed Deep Learning Architecture

The CNN-LSTM model that has been presented will be discussed in this section. Fig. 2 illustrates the CNN-LSTM model that has been proposed for use in image captioning. The diagram depicts the various components of the model. The model is made up of several different layers. The CNN layer is used to first extract characteristics from the image. The CNN layer acquires the knowledge necessary to extract features from images, including edges, shapes, and colors, that are crucial to the process of image captioning. After the output of the CNN layer has been formed into a series of vectors, the next layer is the output of the CNN layer. Each vector represents a different part of the image. The sequence of vectors is then passed to the embedding layer. Each vector is inserted into a space with a high dimension by using the embedding layer. This enables the LSTM layer to learn more complex correlations between the vectors. Long-term dependencies in the vector sequence are learned by the LSTM layer. This is crucial for image captioning since a caption should make sense and be in line with the picture's content. The dropout layer receives the LSTM layer's output after that. By removing part of the LSTM layer neurons at random, the dropout layer stops overfitting. The model is compelled to pick up stronger characteristics as a result. Next, the output from the dropout layer is combined with the output from the layer before it. The purpose of doing this is to depict the picture in a more sophisticated way. The model's last layer is a thick layer. The image's caption is produced by the thick layer. The algorithm predicts one word at a time to create the caption. Long Short-Term Memory is an architecture for recurrent neural networks (RNNs)[26] that is meant to manage and analyze sequences of data, such as time series, natural language, and voice. It is abbreviated as LSTM, which is also the name of the corresponding abbreviation. classic RNNs have difficulty collecting long-term dependencies in data, therefore researchers came up with the idea of LSTMs to solve this problem and overcome the limits of classic RNNs. The capacity of LSTMs to successfully acquire and remember information over extended periods is largely responsible for the explosion in popularity of this type of model. The most important innovation of LSTMs is found in their memory cells, which give them the ability to store information and keep it up to date over time. These memory cells are made up of three gates: the input gate, the forget gate, and the output gate. These are the most important gates. The forget gate selects what information is no longer relevant, the input gate regulates what information is stored in the memory cell, and the output gate decides what information is shown to the network's output. These gates are controlled by three gates: the input gate, the forget gate, and the output gate. This architecture not only allows LSTMs to recognize and recall patterns or relationships within sequential data, but it also helps typical RNNs avoid the problem of vanishing gradients, which is a common issue for these types of networks. DenseNet-201 is a convolutional neural network architecture that was devised by Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger in the year 2016. This model is a modification of the DenseNet-121 architecture, aimed at mitigating the drawbacks commonly observed in conventional deep neural networks like the issue of vanishing gradients and the challenges associated with training very deep networks. Fig. 2 shows the proposed architecture.

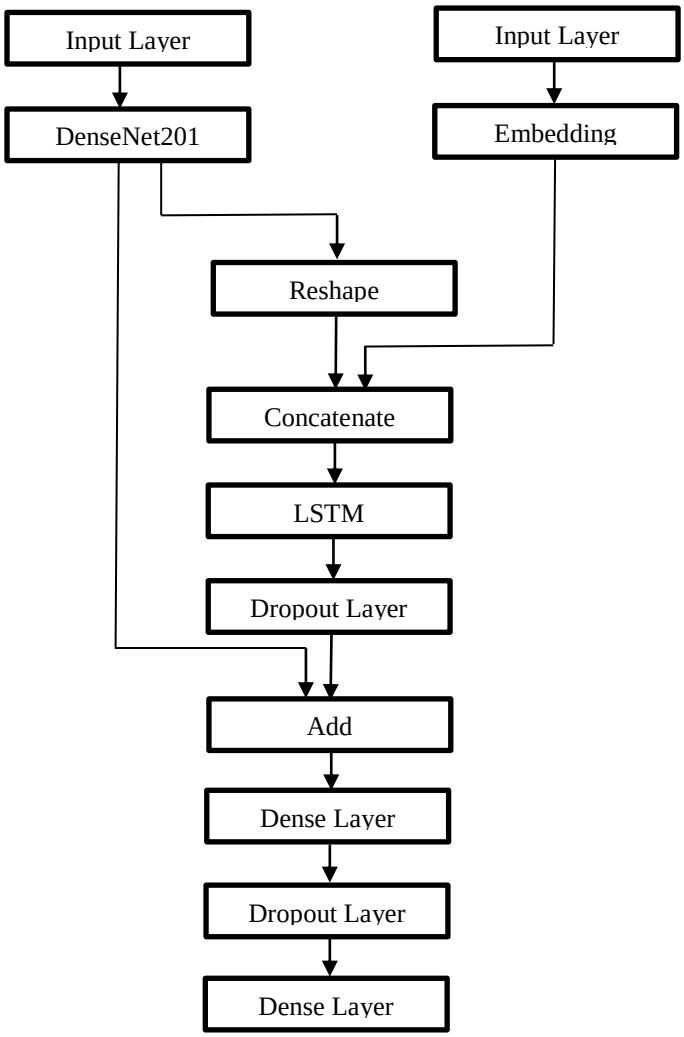

Fig. 2. Proposed model.

IV. Results and Discussion

This section presents a comprehensive account of the outcomes derived from the simulations carried out utilizing the suggested methodology. The dataset utilized in this research was obtained from Kaggle. The dataset (Flickr 8k) was subjected to processing utilizing the suggested technique. The dataset comprises 8,000 photos, each accompanied by five distinct captions. These captions aim to offer comprehensive descriptions of the prominent things and events depicted in the images. The selection of photographs was derived from six distinct Flickr groups, predominantly devoid of prominent individuals or recognizable landmarks. Fig. 3 shows some sample images and their text captions.

Tokenization is a key approach in the field of natural language processing (NLP) that entails the segmentation of text or language into smaller components known as tokens. The tokens employed in this context generally consist of words, sentences, or even individual characters, depending on the task

and the level of detail desired. Tokenization plays a pivotal role as an essential pre-processing step in numerous natural language processing (NLP) applications, facilitating the comprehension and manipulation of human language by computers. In the domain of English language processing, tokenization conventionally entails the division of words by means of spaces or punctuation marks. However, for languages without distinct word delimiters, the process of tokenization may exhibit greater intricacy. The process of tokenization holds significant importance since it facilitates the study of textual data, including many tasks such as text categorization, sentiment analysis, machine translation, and information retrieval. The utilization of algorithms enables the processing of structured and quantified linguistic data, facilitating the extraction of significant insights from textual information, and facilitating effective communication between humans and machines.

The process of image feature extraction in the field of computer vision entails the conversion of unprocessed picture data into a collection of numerical or symbolic descriptors, commonly referred to as features. These features enable more efficient processing and analysis by machine learning algorithms. The process of feature extraction is of utmost importance as it serves to streamline the intricacies associated with visual data, while preserving the vital information required for a multitude of computer vision applications, including but not limited to object recognition, image classification[22], and image retrieval. The techniques employed for feature extraction exhibit a wide range of complexity, encompassing rudimentary approaches such as color histograms and edge detection, as well as sophisticated methods like convolutional neural networks (CNNs) that possess the ability to autonomously acquire pertinent features from images. The collected characteristics play a fundamental role in picture comprehension and facilitate the development of models that possess the ability to identify objects and detect patterns within images. In brief, the process of image feature extraction serves to connect unprocessed visual input with machine learning algorithms, hence enabling the comprehension and examination of images across several domains, including but not limited to autonomous driving and medical imaging.

The notions of training loss and validation loss hold significant importance in the training and evaluation of machine learning models, specifically in the domain of supervised learning tasks like classification and regression. These metrics serve as important indicators for evaluating the performance of the model both during the training process and in subsequent assessments. The training loss is a metric used to evaluate the performance of a machine learning model on the training dataset. Fig. 4. depicts the training and testing loss of the proposed model. The quantification of the inaccuracy or disparity between the predictions made by a model and the actual target values in the training data is referred to as the evaluation of model performance. Throughout the training procedure, the model iteratively modifies its parameters, such as weights and biases, to minimize the loss function. A decrease in training loss signifies that the model is progressively improving its ability to appropriately match the training data. Nevertheless, it is imperative to acknowledge that an excessively low training loss does not guarantee that a model would exhibit strong generalization capabilities when presented with unseen data. Indeed, the phenomenon of obtaining a significantly reduced training loss while exhibiting poor generalization performance serves as an indication of overfitting. Overfitting occurs when a model has efficiently memorized the training dataset yet lacks the ability to accurately predict outcomes for novel, unseen data instances.

Fig. 3. Sample images and captions of flickr 8k dataset.

The validation loss, also known as the test loss, is a metric that evaluates the performance of a model on unseen data, which was not used for training. The validation dataset, which comprises unseen data, serves the purpose of evaluating the model's capacity to generalize, and is separate from the training dataset. The calculation of the validation loss is performed similarly to that of the training loss, where the model's predictions are compared against the actual target values. A low validation loss indicates that the model is effectively generalizing its learned patterns to previously unseen data. The metric functions as a significant determinant of a model's efficacy in forecasting and its capacity to generate precise predictions is when applied to the real-world dataset. The results of the proposed model are depicted in Fig. 5.

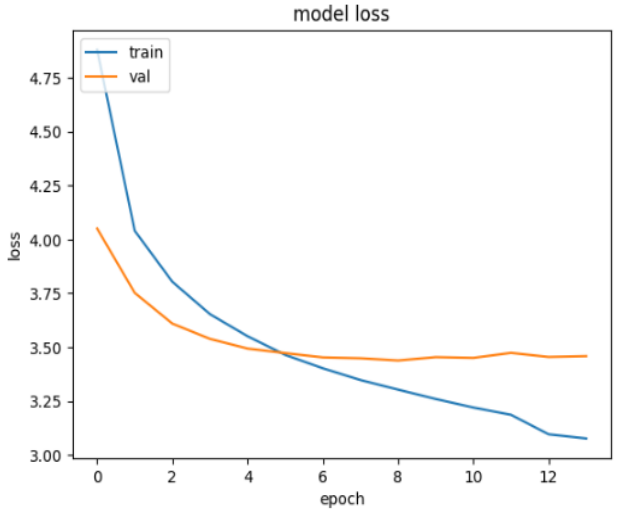

Fig. 4. Training and test loss of the model.

Fig. 5. Captions generated by proposed model on sample test images.

The BLEU metric [23], known as Bilingual Evaluation Understudy, is commonly employed to assess the caliber of machine-generated text, particularly within the domain of machine translation. The metric quantifies the degree of similarity between the text generated by the machine and a reference text, yielding a numerical score that assesses the alignment of the machine-generated text with the reference text produced by humans. The BLEU metric is commonly employed in the fields of natural language processing and machine translation for evaluating the effectiveness of language generating models. The BLEU score operates by doing a comparison between the n-grams, which are consecutive sequences of n words or characters, present in the machine-generated text and those found in the reference text. The evaluation metric evaluates the precision, which quantifies the ratio of n-grams in the text generated by the machine that are also present in the reference text. Additionally, it considers brevity, which takes into consideration the length of the generated text in relation to the reference text. The BLEU score is quantified as a numerical value ranging from 0 to 1, where a higher score signifies a stronger correspondence between the generated text and the reference text. The BLEU score obtained by proposed method is 0.6052. The BLEU metric of the proposed model is compared with existing models and corresponding results are reported in Table I. The Table I compares models' image caption generation performance using the BLEU score, a typical criterion for machine-generated text quality. These models establish image descriptions using CNN and LSTM combinations. The findings show different performance levels: VGG16, a common CNN architecture, scored 0.56 in BLEU, whereas DensNet+LSTM scored 0.57. The Conventional CNN+LSTM model scored 0.39. The Proposed CNN+LSTM have the greatest BLEU score of 0.60, showing its capacity to create picture captions that resemble human-generated reference captions.

TABLE I. BLEU COMPARISON RESULTS

S.No	Model	BLEU
1	VGG16 [19]	0.56
2	DensNet+LSTM [20]	0.57
3	Conventional CNN+LSTM [21]	0.39
4	Proposed CNN+LSTM	0.60

V. CONCLUSION

In conclusion, this research has showcased the capabilities of deep learning methodologies in generating image captions for video summarization. By utilizing the DenseNet201 architecture for extracting image features and deploying GloVe LSTM models for text processing, the proposed model has effectively developed a framework that effectively connects visual and textual content, providing a comprehensive solution for applications related to video summarization. The captions provided offer significant contextual information and important perspectives for video content, hence enhancing its accessibility and interpretability. The proposed framework obtained a BLEU score of 0.60. The image caption model built using the Flickr8k dataset with proposed architecture has limitations stemming from the dataset's relatively small size,

potentially leading to overfitting and a lack of generalization to diverse images. Access to larger and more diverse datasets, possibly incorporating more specialized datasets for nuanced image understanding, will enhance the model's generalization capabilities.

REFERENCES

[1] Poongodi, M., Mounir Hamdi, and Huihui Wang. "Image and audio caps: automated captioning of background sounds and images using deep learning." *Multimedia Systems* (2022): 1-9.

[2] Dognin, Pierre, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. "Image captioning as an assistive technology: lessons learned from VizWiz 2020 challenge." *Journal of Artificial Intelligence Research* 73 (2022): 437-459.

[3] Apostolidis, Evlampios, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. "Video summarization using deep neural networks: A survey." *Proceedings of the IEEE* 109, no. 11 (2021): 1838-1863.

[4] Hussain, Tanveer, Khan Muhammad, Weiping Ding, Jaime Lloret, Sung Wook Baik, and Victor Hugo C. de Albuquerque. "A comprehensive survey of multi-view video summarization." *Pattern Recognition* 109 (2021): 107567.

[5] Behrens, Ronny, Natasha Zhang Foutz, Michael Franklin, Jannis Funk, Fernanda Gutierrez-Navratil, Julian Hofmann, and Ulrike Leibfried. "Leveraging analytics to produce compelling and profitable film content." *Journal of Cultural Economics* 45 (2021): 171-211.

[6] Fajtl, Jiri, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. "Summarizing videos with attention." In *Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14*, pp. 39-54. Springer International Publishing, 2019.

[7] Dilawari, Aniqa, and Muhammad Usman Ghani Khan. "ASoVS: abstractive summarization of video sequences." *IEEE Access* 7 (2019): 29253-29263.

[8] Tiwari, Vasudha, and Charul Bhatnagar. "A survey of recent work on video summarization: approaches and techniques." *Multimedia Tools and Applications* 80, no. 18 (2021): 27187-27221.

[9] Ul Haq, Hafiz Burhan, Muhammad Asif, Maaz Bin Ahmad, Rehan Ashraf, and Toqeer Mahmood. "An effective video summarization framework based on the object of interest using deep learning." *Mathematical Problems in Engineering* 2022 (2022).

[10] Sridevi, M., and Mayuri Kharde. "Video summarization using highlight detection and pairwise deep ranking model." *Procedia Computer Science* 167 (2020): 1839-1848.

[11] Issa, Obada, and Tamer Shanableh. "Static Video Summarization Using Video Coding Features with Frame-Level Temporal Subsampling and Deep Learning." *Applied Sciences* 13, no. 10 (2023): 6065.

[12] Wang, Xu, Yujie Li, Haoyu Wang, Longzhao Huang, and Shuxue Ding. "A Video Summarization Model Based on Deep Reinforcement Learning with Long-Term Dependency." *Sensors* 22, no. 19 (2022): 7689.

[13] Rhevanth, M., Rashad Ahmed, Vithik Shah, and Biju R. Mohan. "Deep Learning Framework based on audio–visual features for video summarization." In *Advanced Machine Intelligence and Signal Processing*, pp. 229-243. Singapore: Springer Nature Singapore, 2022.

[14] Khan, Gulraiz, Saira Jabeen, Muhammad Zeeshan Khan, Muhammad Usman Ghani Khan, and Razi Iqbal. "Blockchain-enabled deep semantic video-to-video summarization for IoT devices." *Computers & Electrical Engineering* 81 (2020): 106524.

[15] Chen, Xiaoning, Mingyang Ma, Runfeng Yang, and Yong Peng. "Multiview feature co-factorization based dictionary selection for video summarization." *IET Image Processing* (2023).

[16] Zheng, Ke, and Xiangdi Chen. "Research on video summarization method based on convolutional neural network." In *International Conference on Neural Networks, Information, and Communication Engineering (NNICE)*, vol. 12258, pp. 52-56. SPIE, 2022.

[17] Balamurugan, G., and J. Jayabharathy. "An integrated framework for abnormal event detection and video summarization using deep learning." (2022).

[18] Sah, Ramesh Kumar. "Video Summarization using Spatio-Temporal Features by Detecting Representative Content based on Supervised Deep Learning." PhD diss., Pulchowk Campus, 2021.

[19] Sri Neha, V., B. Nikhila, K. Deepika, and T. Subetha. "A Comparative Analysis on Image Caption Generator Using Deep Learning Architecture—ResNet and VGG16." In *Computational Vision and Bio-Inspired Computing: Proceedings of ICCVBIC 2021*, pp. 209-218. Singapore: Springer Singapore, 2022.

[20] Deng, Zhenrong, Zhouqin Jiang, Rushi Lan, Wenming Huang, and Xiaonan Luo. "Image captioning using DenseNet network and adaptive attention." *Signal Processing: Image Communication* 85 (2020): 115836.

[21] Das, Ringki, and Thoudam Doren Singh. "Assamese news image caption generation using attention mechanism." *Multimedia Tools and Applications* 81, no. 7 (2022): 10051-10069.

[22] Inayathulla, M., Karthikeyan, C. (2022). Supervised Deep Learning Approach for Generating Dynamic Summary of the Video. In: Suma, V., Baig, Z., Kolandapalayam Shanmugam, S., Lorenz, P. (eds) Inventive Systems and Control. Lecture Notes in Networks and Systems, vol 436. Springer, Singapore. https://doi.org/10.1007/978-981-19-1012-8_18.

[23] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2002.

[24] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077-6086.

[25] J. Wu, T. Chen, H. Wu, Z. Yang, G. Luo and L. Lin, "Fine-Grained Image Captioning With Global-Local Discriminative Objective," in IEEE Transactions on Multimedia, vol. 23,2021, pp. 2413-2427.

[26] Marc Tanti, Albert Gatt, Kenneth P. Camilleri, "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?",Proceedings of the 10th International Conference on Natural Language Generation,2017, Pages 51-60.

[27] Amirian, S., Rasheed, K., Taha, T.R., Arabnia, H.R. (2021). "Automatic Generation of Descriptive Titles for Video Clips Using Deep Learning" In: Advances in Artificial Intelligence and Applied Cognitive Computing. Transactions on Computational Science and Computational Intelligence. Springer, Cham. https://doi.org/10.1007/978-3-030-70296-0_2.

[28] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", Proceedings of the 32nd International Conference on Machine Learning, PMLR 2015,37:2048-2057.