INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Novel Transfer Learning Attitude for Automatic Video Captioning Using Deep Learning Models

J Vaishnavi[1]*, V Narmatha[2]

[1] Research scholar, Department of Computer & Information science, Annamalai University, Annamalai Nagar, Tamil Nadu, India
[2] Assiatant Professor, Department of Computer & Information science, Annamalai University, Annamalai Nagar, Tamil Nadu, India

OPEN ACCESS

Received: 13-09-2022
Accepted: 08-10-2022
Published: 17-11-2022

Citation: Vaishnavi J, Narmatha V (2022) Novel Transfer Learning Attitude for Automatic Video Captioning Using Deep Learning Models. Indian Journal of Science and Technology 15(43): 2325-2335. https://doi.org/10.17485/IJST/v15i43.1846

* **Corresponding author**.

vaishnavijayabal22@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To generate the captions for the videos with less time complexity and high accuracy and also to create captions for each input video frame with particular timestamps. It will be utilized in the crime branch and hearing-impaired people will learn about the happenings of the video fruitfully. **Methods:** The proposed approach experiments with Transfer learning techniques. Modified Inception v3 and Resnet 50 networks are designed to compare the results. The standard MSVD Dataset is utilized to demonstrate the architectures. The performances are compared with the standard performance metrics. **Findings:** The inception v3 model works better than the Resnet 50 architecture for video captioning tasks. It provides the best accuracy at 99.83% with captions for the given input videos than Resnet 50 model. The MSVD dataset is more suitable for the demonstration of the video captioning task. **Novelty:** The two proposed models are modified based on the working of the video captioning tasks. The aggregation of some layers boosts the performance of the models more than ordinary models.

Keywords: Artificial Intelligence; Automatic Captioning; Transfer Learning; Frames; Inception V3; Residual Network50 Model

1 Introduction

The deep learning approach paves the new idea for the automatic generation of subtitles while displaying any video or live telecast by construction models based on the standard design. Deep learning approaches are trying to face the challenges and provide the solution for the captioning problems by implementing pre-trained models with training and testing phases for the given data set. The model is designed as a fusion-based approach for video captioning tasks. The fusion of transfer learning and recurrent neural network to analyze and examine the sentences' correspondence with the visuals of the input video and also to enhance the granularity. The aggregation of the recurrent neural network is mainly performed for the better analysis of text and also for the complete linguistic process[1].

Image captioning is at the forefront of captioning videos. The current researchers are involved in the area of captioning videos by using various deep learning techniques such as convolutional neural networks, long short-term memory, recurrent neural network, generative adversarial network, etc. Most of the techniques are overlapped[2] between image captioning and video captioning tasks. Most video captioning models follow the encoder-decoder structure. CNN and RNN are utilized as the encoder for extracting various features such as visual features, motion features, audio features, etc. long short-term memory, transformer, and generative adversarial networks are employed as the decoder for generating captions. Convolutional networks are popularly used for image-related tasks and also play a vital role in object identification and segmentation etc. it also works better for extracting visual features from the video frames in video captioning tasks. RNN is the unavoidable network that is mainly utilized for text-processing tasks. The extended version of RNN leads to introducing GRU gated recurrent unit network and long short-term memory network for generating captions for the videos. Both CNN and RNN can act as the encoder as well as a decoder in video captioning tasks. The various transfer learning architectures are also integrated into the design for extraction features and also to generate captions such as Alexnet, VGGnet, and Resnet with varying layers. The combination of the above techniques is also experimented with by various researchers for the task of video captioning. The technical platforms utilized to implement the above techniques are Tensor flow, PyTorch, and Keras.

The model of CNN- RNN is designed along with the beam search strategy and greedy search approach separately. The performances are compared and the grammatical issues are overcome by setting the beam search value as 3. The corrected grammatical issues include spelling mistakes, mismatched pronouns, prepositions, nouns, etc. the performances are evaluated and the demonstration is performed by utilizing the benchmark dataset MSVD[3]. The current world is habitted by the usage of more videos in their daily life. This makes the processing of videos more complex. The three different efficient models are equipped for processing the videos with innovative feature extraction techniques. The extracted features play a major role in the final output. The subtitles, visual features, and audio features are the features extracted. The three popular techniques are fused to provide better results such as optical character recognition, object deep learning identification techniques, and automatic speech recognition techniques. Three separate models are designed and implemented in sparks such as faster R-CNN ResNet, Faster R-CNN Inception ResNet V2, and Single Shot Detector MobileNet V2[4].

The video captioning tasks is experimented with various deep learning techniques as both encoder and decoder and evaluated by using the standard performance metrics. The encoder-decoder model is the basic structure for extraction features and generating captions for the videos. Machine learning and deep learning techniques are employed in different structures. Deep learning techniques perform better and provide the highest accuracy due to their high computing capacity, less time-consuming capacity, and processing of a large amount of data. The benchmark datasets such as MSVD, MSR-VTT, Flickr 30, and youtube2text video corpus datasets ate discussed. The evolution of techniques for video captioning tasks is discussed and employed in various models such as template-based models and deep learning-based models. Deep learning approaches are generating good results using the encoder-decoder framework for producing appropriate language descriptions for video scenarios[5]. The attention models are employed by various researchers for video captioning tasks. The gaussian attention technique replaces the soft attention techniques for captioning videos. The complexity of utilizing a fixed-size video is overcome by the Gaussian attention technique. The parametric Gaussian attention technique is also employed for processing frames of the videos. Two different models are structured and employed for generating captions. Steering captioning is also employed for the hierarchical model. The multi-stream hierarchical model varies from the fixed hierarchical model[6].

The captions for the videos are generated based on the encoder-decoder reconstructor. The model is designed with two major parts such as caption generation and video reconstruction. The convolutional neural network is utilized for the complete lexical caption generation along with the multi-label and multi-instance learning to structure. The video reconstructor part enables to recreate of the raw video by using the captions generated. The video reconstructor part minimized the semantic gap between the visuals and lexical terms[7]. The video captioning tasks is highly functional for visually impaired people to recognize the happenings in the video. The convolutional neural network acts as the encoder for extracting the essential features and the long short-term memory structure acts as the decoder for the generation of actual captions for the videos. The model experiments with the benchmark dataset MSVD. The sentences are generated by muting the videos[8].

The vision transformer is fused with the reinforcement learning technique to generate appropriate captions for the video. The transfer learning techniques as Resnet – 101 and Resnet – 152 are utilized for the extraction of the essential features and then the vision transformer is specially employed to extract the needed visual feature. The long short-term memory is equipped to generate the captions by using the extracted features. The model is demonstrated with the standard Msr – vtt dataset to exhibit better results[9]. Image captioning is the base for introducing video captioning. The caption for the image is generated by extracting features. The pre-trained networks are employed for feature extraction. Convolution features are extracted by utilizing the CNN model and the essential object features are extracted by using the YOLOV4 model. The MSCOCO and flicker datasets are utilized to demonstrate the model[10]. Visual reasoning is utilized in various video-related tasks such as visual question - answering, and visual grounding. The language parsing is performed by structuring the natural language grounding model which composes the binary tree structure and automatic process of visual reasoning[(11)].

POS structure is utilized for the sentence formation of the captions which acts as the syntax representation for the sentence generation. It considers both the visuals that appear in the video and also the syntax formation to produce the appropriate captions. The syntactic structure plays a vital role in converting the visual clues into the lexical which describes the visuals. The above two parts are modeled and work in the end–to–end manner. Sequence learning methods are used widely for video description-related tasks for better language description. Part-of-speech models are employed to boost video captioning tasks. POS sequence features help to predict the hidden features[(12)]. The different kinds of representations from the input videos such as motion and Content features are fused as the encoder by constructing the cross-gating block. Here the captions are generated by extracting and fusing various features from the input video and generating captions using the pos structure. The global syntactic structure is envisaged based on the pos generator depending on the fused features. The decoder works based on utilizing a gating strategy by involving global syntactic information to generate accurate captions[(13)].

Visual captioning uses neural models, gaining popularity latently implementing attention mechanisms for the automatic generation of captions in video clippings by focusing on the image subcategories[(14)]. Most approaches for video captioning provides good results with sound background details but in most cases, the missing details are speed breakers for sentence generation. The joint vision language method can retrieve captions for images or frames in the video clippings similar to the encoder-decoder model[(15)] but the captions are received from the embedding region. This model copies the captions through a pointer network generator for exclusive video clippings. Another research methodology involves a hierarchical model with a multi-stream boundary model[(16)] using steered captioning method with parameters to locate the areas in the video. Intrinsic features identify temporal structures at specific time slots to narrate the video.

The existing models are trained and executed with complex structures with high consumption of time. The proposed models are designed with the simple architecture by the aggregation of the dense layers for extracting the essential features and also the sigmoid activation function for predicting the frames with the actual captions with less time consumption and less computational cost. This modification boosts the performance of the architectures and also provides better accuracy. The pre-trained networks are suitable for many problems based on the nature, complexity, and process of the problem. As per the nature of the problem and the layers of the architecture, the modified Inception V3 model provides better results than the modified Resnet 50 model.

2 Methodology

2.1 Role and Outline of the Model

The proposed work deals with the automatic generation of subtitles while displaying a live scene or some recorded work. The data set is collected online from sources like MSVD and preprocessed. Some sample video scenes are shown in Figure 1. In this stage, the videos are converted into frames and the frames are rescaled to standard size. After preprocessing the frames are classified based on the features extracted and two models are suggested. One is Inception and the other is the Resnet model. Finally, the frames are classified into 50 categories.

Fig 1. Sample video extracts

The organization of this paper first forecast the advantages of video captioning. The next section deals with the proposed methodology where the videos are preprocessed and classified. The final section deals with experimental results and the conclusion.

2.2 Need for Video Captioning

A leaping step in technology development is the automatic generation of captions suitable for video clippings made possible by numerous algorithms and models. There are numerous advantages of video captioning and some of them are summarized. They are as follows,

- Accessible for physically challenged persons who have a problem with hearing and deaf people. A person can watch the video clippings and understand the contents of the video by using the captions displayed along with the scenes. The captions are generated in time combination so that the viewers do not miss any vital content.
- To avoid certain law problems in overseas countries, particular videos should be played along with captions so they can be easily acceptable.
- Highly useful for online education and training programs as the captions improve comprehension by combining both the audio and scripts read, and memory will be stronger. Captions help non-native people who are unable to understand the accent or slang of the language.
- Suitable for environments where audio or sound is not allowed. In that case, captions only help the viewers to understand the scenes of the video and the contents of the image. Captions are also needed in mute situations like temples, travel, college, or in the office where sound is prohibited.
- An improved learning experience is provided by the transcripts or captions where the viewers can search for keywords. The viewers can traverse the entire script using the keywords for a better customer experience.
- Creating subtitles using captions format to meet the need of customers belonging to foreign languages. Converted captions are useful for non-English speaking regions so that content is reachable to all. Sample video frames with captions are displayed in Figure 2.

Vet Y: Dogs are designed to walk flatland
- their muscle and joint structures are developed
to hunt in vast plains,
Vet Y: Dogs are designed to walk flatland
- their muscle and joint structures are developed
to hunt in vast plains,

Fig 2. Sample videos with subtitles

2.3 Proposed Approach

The main aim of this work is to produce automatic captions while playing a video clip on entertainment or other educational channels. Deep learning-based pre-trained models accomplish this task by providing better results. The outline for this model is given in Figure 3.

The proposed method uses two models namely Inception V3 and Residual network 50 models for the classification of the images once they are pre-processed. Before this process, the collected data set videos are converted into frames. The frames are classified into different actions which are pre-loaded in the models. So the models contain both training and testing phases where the data set contains both videos and captions for the videos

2.4 Preprocessing

In this section, the videos in the data set are pre-processed to give proper structure to the frames in the video before the process of feature extraction. The quality can be enriched to avoid unnecessary biases and sharpen needed features for processing. Pre-processing is the step required to clean the images in the video to give input to the classifier. Especially deep learning models like CNN architecture take the same sized images as input of standard scale.

There are two pre-processing steps in this method implemented for video captioning. They are first converting the videos into frames and next, rescaling the converted frames of various sizes to the standard size of 200 x 200 pixels. The two approaches are explained with a clear picture as follows.

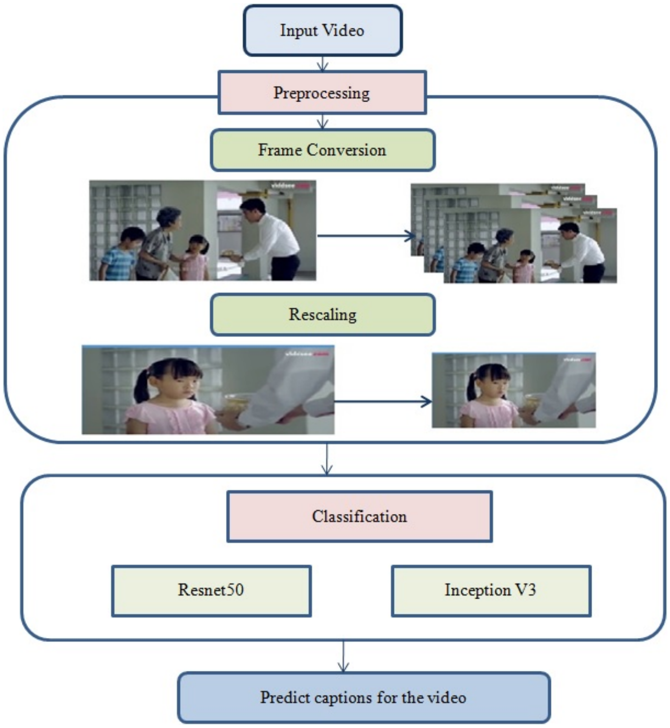


Fig 3. Block diagram of the proposed model

2.4.1 Conversion of videos into frames

For the automatic generation of captions in the video show or live relay, the video must be first converted into frames. Then the frame must be identified and the actions they are performing should be described. Because videos are consists of various scenes with several frames which are organized in a particular order. The images are stored based on the classes.

For converting the video into frames, an algorithm is developed suitable for online shows and educational channels. The steps are,

- The first step is uploading the video by choosing the stored file already or by importing it from the camera. The file can be recovered from the Google Drive account also.
- The next step is preparing a sequence of images from the video by selecting the speed of the frame to determine the time to break. The speed types are fast, slow, and medium ranging from 0 to 1.5 seconds.
- The time decides the frequency of how many frames are to be captured. After the frame is recorded the format to be saved is chosen. There are many extensions like MOV, MP4, etc.
- The next important step is to activate the stop motion effect where each frame is separated by numerous time partitions to generate an order of frames that create a particular action.
- Once the frames are saved, they can be checked for accuracy. If satisfying the frames can be stored otherwise the frames can be edited to improve the resolution of the images in the frame.

By this method, the videos collected from the dataset are converted into frames for processing. The picture below describes the first step in pre-processing video conversion into frames as Figure 4.

2.4.2 Frame rescaling

The next step in pre-processing is resizing the converted frames into an equal or standard size like 200 x 200 pixels in each dimension. The videos in the data set are converted into frames in the previous step. Once the videos are transformed, they are used for classification. In the next preprocessing step, the frames are resized to a standard size of 200 x 200 which is suitable for further action. For changing the size of the frame which is to increase or decrease the pixels, the method used is interpolation which is applied to all the frames. The method used for reducing the image is the pixel area method called the inter-area method

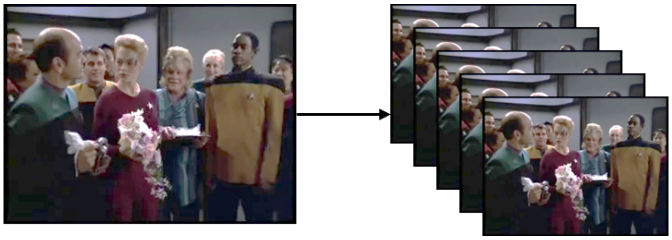


Fig 4. Conversion of video into frames.

where the different size frames are changed to the standard size. Using this method, the frames are converted into a regular size which is used for classification. The figure below shows the original image and the rescaled one as Figure 5.

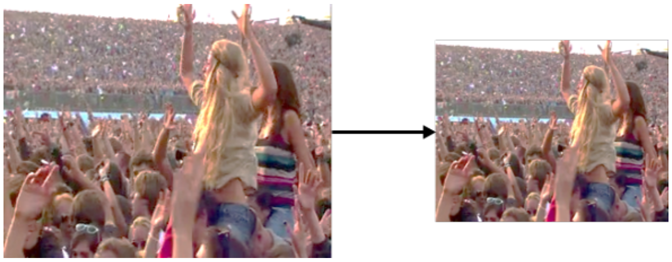


Fig 5. Rescaling of frames into a standard size.

2.5 Classification

After pre-processing step, the frames are classified using the pre-trained models into several actions. The two suitable prototypes used in this work are Inception V3 and Residual Network 50 model. Using these two methods the given video frames from the data set are classified according to the pre-loaded messages. The accuracy is determined by the fact that the percentage of matching subtitles to the scenes is more in either of the two models. The two architectures are clearly explained in the forthcoming sections.

2.5.1 Inception v3 model

The famous classification method used for image identification is the Inception V3 model, a wide network referred to as the Image Recognition model with 48-layer network architecture. The model is built with the fundamental element called the Inception Cell where convolution operations are executed at various levels and the average is taken as the final output. For an enhanced understanding of this design, the basic concepts are explained. The model is constructed by various segments. They are,

- The first part contains filters with various shapes used in parallel to arrest all the information from the images. Usually, three convolution layers with filter sizes (3x 3, 5x5and 1x1) are used with one pooling layer.
- The second section deals with dimension reduction by decreasing the parameter strength to remove the overfitting problem.
- Next 5x5 convolutions after 1x1 and 3x3 are achieved to reduce the computation resources.
- Finally, Auxiliary classifiers are used to prevent the gradient problem that may disappear in the wider network

The Inception V3 architecture contains important building blocks which constitute the framework of the ideal. They are Block A, B, and C to perform the calculations for deriving the features from the images. The block functions are,

Block A - Factorization step that is small convolutions replace big convolution layer for reducing the parameters. A single 5x5 convolution layer is exchanged by two 3x3 convolutions.

Block B - Factorization into asymmetric convolutions reducing more parameters. Single 3x3 convolutions are swapped by 1x3 and 3x1 layers decreasing 33 percent of parameters.

Block C - Factorization of standard 7x7 convolution into asymmetric 1x7 and 7x1 convolution and replace with a smaller convolution layer. Again, 7x7 convolutions are replaced by 3x3 convolutions.

Using these building blocks, the basic architecture is constructed. The architecture is given in Figure 6.

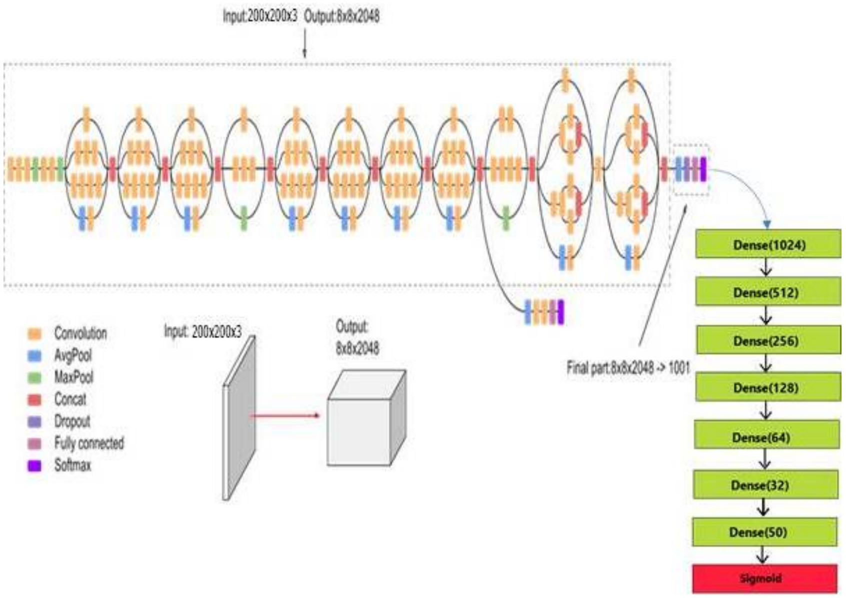


Fig 6. Modified Inception V3 architecture

2.5.2 The Modified Architecture of the Inception v3 model

The standard architecture for the Inception V3 model is used for the classification of given frames with slight modifications to be suitable for processing the video frames. The outline used for classification consists of fundamental steps revised as follows,

1. First, the image to be given as input is acquired after pre-processing, and usually, the size of the image by default is 299x299x3.
2. The input image is processed by 3 convolution layers with filter size 3x3and the number of filters used is 32 for the first two layers and 64 for the third layer with stride 2.
3. Three max-pooling layers are used in between convolutional layers with window size 2x2 and stride 2.
4. The activation function used in the processing of frames is the Relu linear function to provide input only if it is positive.
5. Flatten layer is used to convert the multidimensional input values to dimensional value that represents the class the frame belongs to. In this model, the value can be from 1 to 50 representing the captions the frame can display.
6. The linear value is fed to the dense layer which classifies the input image into one of the fifty given classes.
7. The final layer is the sigmoid classifier that is used for the classification of the video frames into the specified classes that give the captions matching the frames.

The proposed architecture described above is used for classifying the video frames from the benchmark data set into predefined captions. The performance measures are calculated and compared with the other model in the next section.

2.5.3 Residual Network50 model

Another distinguished model for the classification of images based on CNN architecture is the Residual Neural Network model used to solve many problems created in deep networks. The resnet model provides the solution for the vanishing gradient problem by providing skip connections where the input value combines with the function value to give the output by skipping some layers so that small values are avoided. The formula for the Resnet model is given in equation (1).

$$y = F(x) + x \tag{1}$$

Where x is the input and F(x) is the function of x and y is the output value.

The proposed architecture for the Resnet 50 model is constructed by making some vital alterations to the basic model and shown in Figure 7 explained clearly as follows.

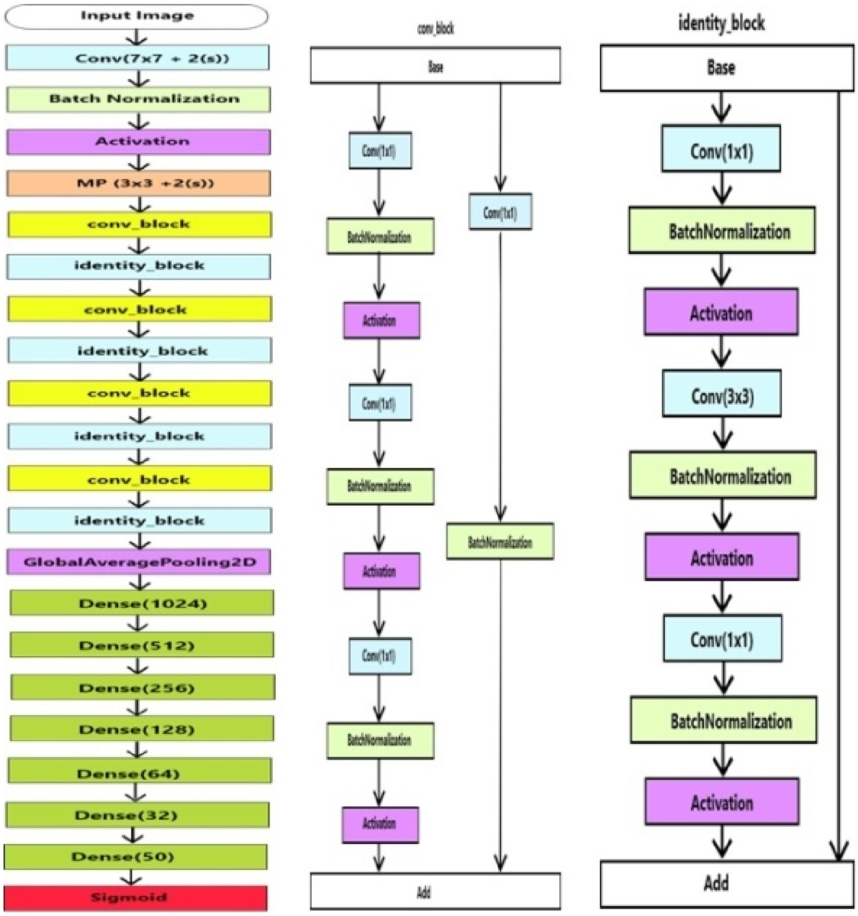


Fig 7. Modified Resnet50 architecture

2.5.4 The Modified Architecture of the Resnet50 model

The basic Residual network model consists of 50 layers as the name implies, and is the deeper network for training complex data for better enactment. The network is assembled by sharing the weights in a shallow network because of their performance and the same logic is used to construct deeper networks and identity mapping is present in the gap. The basic building block of a Residual network is a group of blocks with convolution and max pooling layers. Activation functions and batch normalization functions are used to get the proper result. Skip connections allow the gradient to be directly back propagated to earlier layers.

The basic model contains 2 functions namely a convolution block and an identity block executed based on the fact whether the input and output size of the image are equal. If two values are the same Identity block is performed else a Convolution block is implemented. For the classification of video frames, the basic model is modified suitable for the current problem which is the novel approach of adding hidden layers to give 50 classes. The sketch of the model is described as follows,

1. After the input image is retrieved first convolution layer is performed with a 7x7 filter size and 64 filters with stride 2 are used which is predefined for this model.
2. After this, Batch Normalization is performed and Activation is accomplished using the relu function to avoid the vanishing gradient problem.
3. After convolution layer 3x3 max pooling layer is performed with stride 2.
4. The next 4 convolution blocks and Identity blocks are performed based on the size of the input and output value.

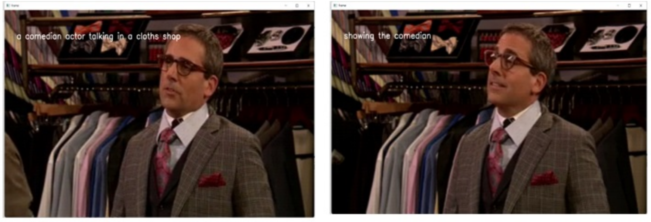
5. In the Identity block, the previous layer value is taken as the base, and a convolution layer with 1x1 filters is performed.
6. In the convolution block, 1x 1 convolution is performed on the base value from the previous layer, and an additional 1x1 convolution is performed.
7. Finally, the global average pooling function is performed to get the feature values.
8. Then dense layers are added and the final value is given to the sigmoid classifier for the classification of video frames into desirable classes.

Flattening and appending the dense layers provide the number of classes and finally, the features are reduced and the last dense layer receives the important features to identify the class the frames belong to.

Using 2 classification models the video frames are categorized into 50 classes where each class represents unique actions. The results are compared for accuracy.

2.6 Performance Comparison of Classification Models

The video frames from the data set are pre-processed and classified using two CNN architecture-based models namely the Inception V3 model and the Residual Network 50 model. The output from the two models is displayed and compared for accuracy. The result video frames with captions are given. The result of the Inception model is given Figure 8a and Resnet model is given in Figure 8b.



a comedian actor talking in a cloths shop

showing the comedian

Fig 8. A) The result of the Inception V3 model, B) The result of the Resnet50 model

3 Result and discussion

The proposed model is designed with two different modified pre-trained architectures such as Inception V3 and Resnet 50 models. The modification is made by aggregating the dense layers for classifying the frames with appropriate captions and also the sigmoid activation function is also appended for the best results by predicting the appropriate captions for the input frames. The proposed works are highly differentiated from the existing architectures by modifying the layers with less time consumption and less computational cost. The unique structure of the Inception V3 and Resnet 50 models improves the performance of ordinary architecture. Despite employing different architectures separately for feature extraction and classifying frames with captions, the single modified structures work as both feature extractor and classifier for generating appropriate captions. This process makes the model simpler and faster to provide accurate results.

The video frames are classified into 50 different classes and based on these classes the actions performed by the scenes are displayed. The volume of accuracy is calculated using the evaluation metrics and the values are displayed in Table 1.

Table 1. Performance Comparison for Proposed Methods

Performance Metrics	Values For Inception Method In (%)	Values For Residual Network Method In (%)
Accuracy	99.83	81.22
Precision	98.72	78.36
Recall	98.68	74.29
F1 – Score	98.66	71.20

The two proposed methods are evaluated and samples of nearly 4177 images from the data set are selected and processed to produce the result. The accuracy of these two models is depicted in the Bar graph below. The inception model graph is shown in Figure 9a and the Residual model is shown in Figure 9b. Figure 9 explains that the loss is reduced meanwhile the accuracy increases by epochs for both training and testing.

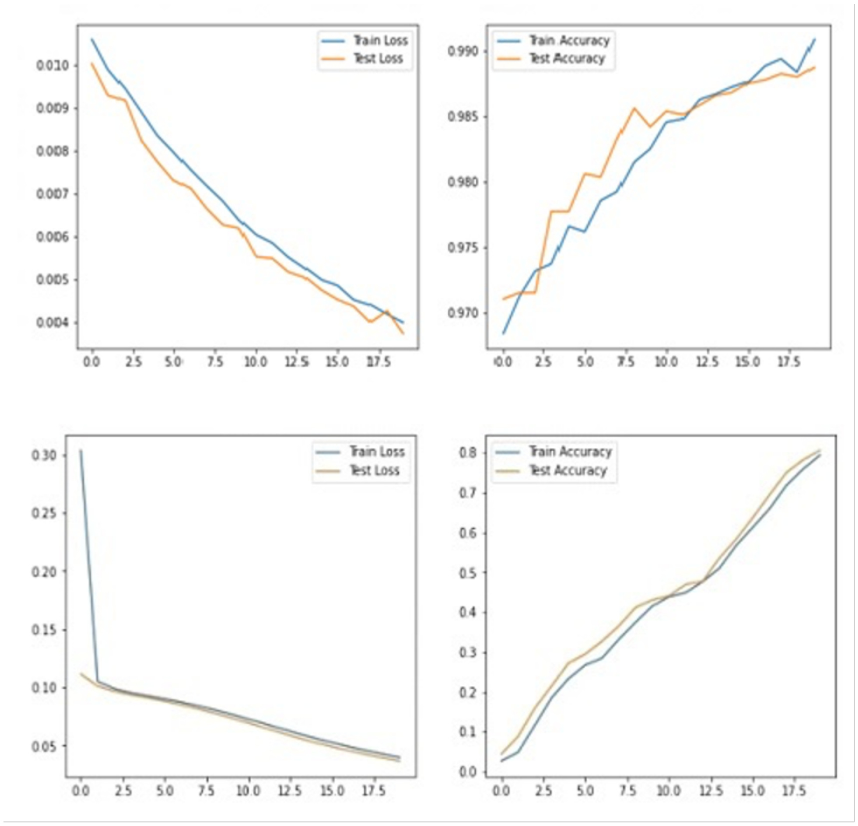


Fig 9. A) Inception Model Graph, B) Resnet 50 Model Graph

4 Conclusion

This study proposed novel classification techniques for the generation of captions spontaneously using Inception V3 and Residual 50 model and the accuracy is checked with the outcoming results. The pretrained networks such as Inception V3 and Resnet 50 are modified by aggregating some dense layers and sigmoid activation functions to boost the performance of the architecture. It leads to providing the highest accuracy of ordinary architecture. The best accuracy 99.83% is obtained from the Inception v3 model. Comparing two proposed modified pre-trained networks, Inception V3 provides better results with appropriate captions than Resnet 50 model. This idea can be stretched for other image processing problems to acquire superlative results compared with the standard model and also it can be extended by employing the fused pre-trained networks for further improvement in the video captioning tasks.

References

1) Sasikala S, Ramesh S, Gomathi S, Balambigai S, Anbumani V. Transfer learning based recurrent neural network algorithm for linguistic analysis. *Concurrency and Computation: Practice and Experience*. 2022;34(5). Available from: https://doi.org/10.1002/cpe.6708.
2) Amirian S, Rasheed K, Taha TR, Arabnia HR. Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap. *IEEE Access*. 2020;8:218386–218400. Available from: https://doi:10.1109/ACCESS.2020.3042484.
3) Padmawar P, Borade R, Hol A. Video Captioning Using Neural Networks. *International Journal for Research in Applied Science and Engineering Technology*. 2022;10(5):1228. Available from: https://doi.org/10.22214/ijraset.2022.42506.
4) Tc, Phan AC, Phan HP, Cao TN, Trieu. Content-Based Video Big Data Retrieval with Extensive Features and Deep Learning. *Applied Sciences*. 2022;12:6753. Available from: https://doi.org/10.3390/app12136753.
5) Amaresh S, Chitrakala. Video Captioning using Deep Learning: An Overview of Methods, Datasets and Metrics. 2019. Available from: https://doi:10.1109/ICCSP.2019.8698097.

6) Samleti S, Mishra A, Jhajhria A, Rai SK, Malik G. Real-Time Video Captioning Using Deep Learning. *International Journal of Engineering Research & Technology (IJERT)*;2021(12):360–366. Available from: https://doi:10.17577/IJERTV10IS120054.

7) Ji W, Wang R. A Multi-instance Multi-label Dual Learning Approach for Video Captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications*. 2021;17(2s):1–18. Available from: https://doi.org/10.1145/3446792.

8) Eg, Özer IN, Karapınar S, Başbuğ S, Turan A, Utku MA, et al. Deep Learning based, a New Model for Video Captioning. *International Journal of Advanced Computer Science and Applications*. 2020;11(3). Available from: https://doi:10.14569/IJACSA.2020.0110365.

9) Zhao H, Chen Z, Guo L, Han Z. Video captioning based on vision transformer and reinforcement learning. *PeerJ Computer Science*;8:e916. Available from: https://doi:10.7717/peerj-cs.916.

10) Malla M, Jafar A, Ghneim N. The image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data*;2022. Available from: https://doi.org/10.1186/s40537-022-00571-w.

11) Hong R, Liu D, Mo X, He X, Zhang H. Learning to Compose and Reason with Language Tree Structures for Visual Grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022;44(2):684–696. Available from: https://doi:10.1109/TPAMI.2019.2911066.

12) Hou J, Wu X, Zhao W, Luo J, Jia Y. Joint Syntax Representation Learning and Visual Cue Translation for Video Captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019. Available from: https:doi:10.1109/ICCV.2019.00901.

13) Wang B, Ma L, Zhang W, Jiang W, Wang J, Liu W. Controllable Video Captioning With POS Sequence Guidance Based on Gated Fusion Network. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019. Available from: https://doi.org/10.48550/arXiv.1908.10072.

14) Ma CY, Kalantidis Y, Alregib G, Vajda P, Rohrbach M, Kira Z. Learning to Generate Grounded Visual Captions Without Localization Supervision, Lecture Notes in Computer Science book series. *LNIP*. 2020;12363. Available from: https://doi.org/10.1007/978-3-030-58523-5_21.

15) Rimle P, Dogan-Schonberger P, Gross M. Enriching Video Captions With Contextual Text. *International Conference on Pattern Recognition (ICPR)*. 2021. Available from: https://doi:10.1109/ICPR48806.2021.9412008.

16) Islam S, Dash A, Seum A, Raj AH, Hossain T, Shah FM. Exploring video captioning techniques: A comprehensive survey on deep learning methods. *SN Computer Science*. 2021;2(2). Available from: https://doi.org/10.1007/s42979-021-00487-x.