Video Labelling Using Deep Learning

Mansi Bhavsar, Prof Avinash Chaudhary, Prof Safvan Vhora

Department of Computer Engineering, Government Engineering College Modasa
Gujarat Technological University

much more functionality so that in this we label the video using deep learning due to the

In this research, we explore the domain of video labeling using the UCF-101 dataset. The objective is to develop an advanced video processing system for action recognition. Leveraging previous research papers as a foundation, our study aims to enhance video labeling techniques by implementing machine learning algorithms capable of accurately identifying actions within the given video input. By utilizing the UCF-101 dataset, we seek to create an intelligent model that not only extracts relevant information from videos but also provides concise and meaningful label of the actions depicted. This research contributes existing knowledge with innovative approaches, like deep learning and other dataset.

Keyword : Feature Selection, Feature Extraction, Convolution Neural Network ,ResNet , 2D+1D CNN

• Introduction

As we know nowadays media is very useful in each field. At present we show different types of media that are used for that we will determine the label of the video using the label of the video we can recommend other media to the user or analyze the video and fact novel malicious codes trade constantly their signatures, static methods are not appropriate to hit upon them. In the last a long time, the creation of machine studying strategies has contributed tremendous price in detecting new malware, due to their generalization capacity. Device studying models are based on tiers: training and prediction. [1]

The training stage is predicated on a training dataset that contains three phases. The primary section includes extracting a large quantity of features from the different videos inside the training dataset. The second segment consists of rejecting non-pertinent features primarily based on appropriate choice strategies. [3] The third segment consists of the use of one or greater classification models that will analyze to distinguish between malicious and benign documents. Those fashions subsequently turn out to be able to provide accurate predictions dealing with new executable documents inside the prediction level. The selection of both appropriate input features and classification versions leads to the improvement of prediction rates in less time. [1]

In this survey, we mainly focus on vision-based action recognition systems that use a video camera as the primary sensor and incorporate video analysis components used to determine the action in the video. [13]

In this research we use the deep learning technique convolutional neural network method which is used for image recognition and image detection using feature selection and feature extraction methods in this research propose a custom CNN model for video labeling.

Literature Review

There are multiple different technologies and datasets which are used for different image and video processing methods with different categories data like sports, abnormal activities [18], on different parameters. Vision based action recognition systems that use video camera as the primary sensor and incorporate video analysis [2]. component used to determine the action in the video[13].Content Based Human action recognition that use content shape and color to identify the content image or video [5].

Most popular action classification datasets, such as Hollywood-2, HMDB51, UCF101, UCF-50 consist of short clips, manually trimmed to capture a single action. These datasets are ideally suited for training fully supervised, whole-clip, forced-choice video classifiers.[12] Recently, datasets, such as Sports1M [4], YouTube-8M [1], Something-something [12], SLAC [48], Moments in Time [9], and Kinetics [4] have focused on large-scale video classification, often with automatically generated – and hence potentially noisy – annotations. They serve a valuable purpose but address a different need than AVA. [12]

Maintaining temporal consistency in video labeling remains a challenge, especially in scenarios with complex and rapid movements.[13] Approaches such as attention mechanisms and spatiotemporal feature aggregation aim to address this issue.The availability of large-scale labeled video datasets is crucial for training deep learning models.[2] However, the creation and annotation of such datasets pose challenges, limiting the development of robust video labeling models.

Objective of Study

The primary objectives of this study are:

- To Investigate CNN in Video Labeling: Examine the effectiveness of Convolutional Neural Networks in the context of video labeling, particularly focusing on their capacity for feature extraction and recognition.
- Utilize the UCF 101 Dataset: Apply the CNN model to the UCF 101 dataset, a diverse collection of human actions in videos, to evaluate the model's performance across various scenarios and action categories.
- Enhance Accuracy and Efficiency: Improve the accuracy and efficiency of video labeling through the implementation of a custom CNN model,

considering feature selection and extraction methods.

- Address Security Challenges: Address the limitations of traditional static methods in the detection of malicious codes within video content by employing dynamic CNN- based approaches.

Scope Of Study

Video Labeling is used in diverse fields Sports Analysis, It investigates the use of advanced algorithms to automatically identify and label key elements in sports event footage, such as players and actions. The goal is to enhance sports analytics efficiency, providing automated insights for coaches and enthusiasts. Automatic Vehicles, Video content recommendation, Virtual Reality, Quality Control , Security surveillance, Object detection , Sentiment Analysis analyze and classify the emotional content of videos, with applications in content creation, marketing, and user experience optimization.

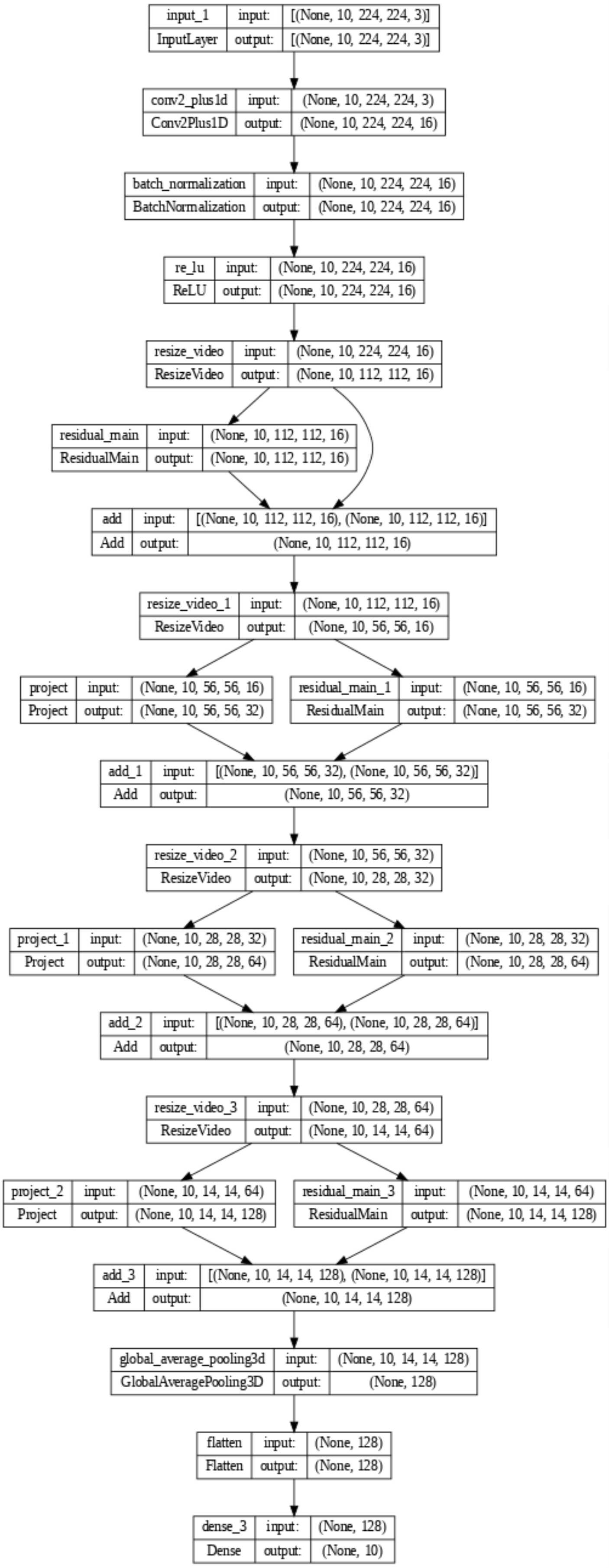
Features Selection

Feature selection involves the identification and prioritization of relevant characteristics Within video frames. In the training phase, a diverse set of features is initially extracted from the video data. However, not all of these features may contribute equally to the task of video labeling. Feature selection methods are employed to filter out non-pertinent features and retain those that are most informative. This ensures that the CNN model focuses on salient information, reducing computational complexity and preventing the inclusion of noise in the training process. By selecting the most discriminative features, the model becomes more adept at recognizing and categorizing actions within videos.

Feature Extraction

The selection of input features is a primary assignment in each machine learning studies. In malware detection field, these features can both be raw records contained within the documents with a purpose to be examined, or the end result of processing raw records. Each benign and malicious documents are considered for the training of the chosen machine learning model. [4]

A critical aspect of the research involves feature extraction techniques. Relevant features are identified and extracted from video frames during the training phase, enhancing the model's ability to discern meaningful patterns. This ensures that the CNN model focuses on salient information, improving both efficiency and accuracy in video labeling.


input_1 | input: | [(None, 10, 224, 224, 3)]
InputLayer | output: | [(None, 10, 224, 224, 3)]
conv2_plus1d | input: | (None, 10, 224, 224, 3)
Conv2Plus1D | output: | (None, 10, 224, 224, 16)
batch_normalization | input: | (None, 10, 224, 224, 16)
BatchNormalization | output: | (None, 10, 224, 224, 16)
re_lu | input: | (None, 10, 224, 224, 16)
ReLU | output: | (None, 10, 224, 224, 16)
resize_video | input: | (None, 10, 224, 224, 16)
ResizeVideo | output: | (None, 10, 112, 112, 16)
residual_main | input: | (None, 10, 112, 112, 16)
ResidualMain | output: | (None, 10, 112, 112, 16)
add | input: | [(None, 10, 112, 112, 16), (None, 10, 112, 112, 16)]
Add | output: | (None, 10, 112, 112, 16)
resize_video_1 | input: | (None, 10, 112, 112, 16)
ResizeVideo | output: | (None, 10, 56, 56, 16)
project | input: | (None, 10, 56, 56, 16)
Project | output: | (None, 10, 56, 56, 32)
residual_main_1 | input: | (None, 10, 56, 56, 16)
ResidualMain | output: | (None, 10, 56, 56, 32)
add_1 | input: | [(None, 10, 56, 56, 32), (None, 10, 56, 56, 32)]
Add | output: | (None, 10, 56, 56, 32)
resize_video_2 | input: | (None, 10, 56, 56, 32)
ResizeVideo | output: | (None, 10, 28, 28, 32)
project_1 | input: | (None, 10, 28, 28, 32)
Project | output: | (None, 10, 28, 28, 64)
residual_main_2 | input: | (None, 10, 28, 28, 32)
ResidualMain | output: | (None, 10, 28, 28, 64)
add_2 | input: | [(None, 10, 28, 28, 64), (None, 10, 28, 28, 64)]
Add | output: | (None, 10, 28, 28, 64)
resize_video_3 | input: | (None, 10, 28, 28, 64)
ResizeVideo | output: | (None, 10, 14, 14, 64)
project_2 | input: | (None, 10, 14, 14, 64)
Project | output: | (None, 10, 14, 14, 128)
residual_main_3 | input: | (None, 10, 14, 14, 64)
ResidualMain | output: | (None, 10, 14, 14, 128)
add_3 | input: | [(None, 10, 14, 14, 128), (None, 10, 14, 14, 128)]
Add | output: | (None, 10, 14, 14, 128)
global_average_pooling3d | input: | (None, 10, 14, 14, 128)
GlobalAveragePooling3D | output: | (None, 128)
flatten | input: | (None, 128)
Flatten | output: | (None, 128)
dense_3 | input: | (None, 128)
Dense | output: | (None, 10)

Proposed Model

- ## 3D Convolutional Neural Network

Convolutional Neural Networks are a class of deep neural networks designed for visual processing tasks. They leverage convolutional layers to automatically and adaptively learn hierarchical features from input data. The key components of a CNN include convolutional layers, pooling layers, and fully connected layers. Convolutional layers apply filters to input data, capturing spatial hierarchies and patterns, while pooling layers down sample the spatial dimensions, reducing computational complexity.

1. Convolutional Layers: Convolutional layers are the foundation of CNNs, responsible for extracting local patterns and features from input data. These layers employ convolutional operations, where small filters or kernels move across the input data, performing element-wise multiplications and aggregating the results. This process allows the network to automatically detect and learn hierarchical representations of features, capturing low-level details in early layers and progressively more complex patterns in deeper layers.

Video labeling, convolutional layers play a crucial role in recognizing spatial structures within individual frames and capturing temporal dependencies across consecutive frames. The filters adaptively learn to identify edges, textures, and other visual cues relevant to the labeled actions in videos.

2. Pooling Layers: Pooling layers are interspersed between convolutional layers and serve to down sample the spatial dimensions of the feature maps, reducing computational complexity while retaining essential information. Max pooling, a common pooling technique, selects the maximum value from a group of neighboring values, effectively emphasizing the most prominent features.

Pooling layers contribute to the model's ability to focus on the most informative parts of each frame. They help in creating a more abstract and compact representation of the learned features, making the network less sensitive to variations in the precise location of those features.

3. Fully Connected Layers: After the convolutional and pooling layers, fully connected layers are often employed to make predictions based on the extracted features. These layers connect every neuron to every neuron in the subsequent layer, allowing the network to learn global dependencies and relationships. In the context of video labeling, fully connected layers integrate the spatial and temporal features learned by the earlier layers to make predictions about the actions or labels associated with the input video sequence.

4. Activation Functions: Activation functions introduce non-linearity into the CNN, enabling the model to learn complex mappings between inputs and outputs. Common activation functions include Rectified Linear Unit (Relook), which replaces negative values with zero, and Sigmoid, used for binary classification tasks. These functions contribute to the network's ability to capture intricate patterns and representations in the data.

5. Training Process: During the training process, the CNN learns to optimize its weights and biases using labeled training data. The back propagation algorithm is employed to minimize the difference between the predicted outputs and the ground truth labels. This iterative process fine-tunes the network's parameters, enabling it to generalize well to unseen data.

6. Transfer Learning: In video labeling research, transfer learning is often utilized. Pre-trained CNN models on large image datasets, such as ImageNet, can be fine-tuned for video-related tasks. This approach leverages the knowledge gained from general visual patterns and allows for effective training on smaller, task-specific datasets like UCF101, enhancing the model's performance.

.

- **Review Table of different methods for Video Labeling**

No.	**Title**	**Dataset**	**Model / Method**	**Proposed Method**	**Accuracy**
1.	A distributed Content-Based Video Retrieval system for large datasets. Journal of Big Data, 8(1), 1-26.[5]	HMDB51	MLT,MHSA	STDHA(Spatial-Temporal Dual-Headed Attention)	78%
2.	Attention-based bidirectional-long shortterm memory for abnormal human activity detection	Custom Dataset	VGGNet,I3D CNN	HAR	81.32%
3.	Video Representation Learning Using Discriminative Pooling[8]	HMDB-51,Charades	CNN	SVMP	63.7%
4.	Video MAE V2: Scaling Video Masked Autoencoders with Dual Masking[11]	Hollywood 2	CNN	Multiple feature extraction	75%

5.	Video-Based Human Activity Recognition Using Deep Learning Approaches. Sensors, 23(14), 6384.	Hollywood 2,UCF50	Feature Extraction , Distance Measure	Rotation forest Classifier	78.63%
6.	Attention-based bidirectional-long short-term memory for abnormal human activity detection. Scientific Reports, 13(1), 14442.	UCF-101 ,HMDB-51	3D Convolutional Autoencoder ,Network Architecture	3D-CAE Networks	81.50%
7.	Sports video classification framework using enhanced threshold based Keyframe Selection algorithm and customized CNN on UCF101 and Sports1-M dataset. Computational Intelligence and Neuroscience, 2022.	UCF101	CNN	CNN	79.75%
8.	Spatiotemporal contrastive video representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6964-6974).[9]	k-600	Encoder, decoder	CVRL framework	72.9%
9.	A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.[14]	UCF101	I3D-ResNet-50	UCF101, HMDB51	72.3%

Proposed Method

Our research focuses on enhancing video labeling for action recognition, utilizing the UCF-101 dataset. We've successfully implemented deep learning algorithms, specifically a Custom 3D Convolutional Neural Network model, to identify actions in videos and assign meaningful labels. This work amalgamates established knowledge with innovative techniques, providing a valuable contribution to the field. Our approach enhances video content analysis and action recognition, applicable across diverse domains such as sports analysis and entertainment.

- **Input Layer:** This is where your video data is fed into the network. It typically consists of multiple frames, forming the temporal dimension of your data.
- **2D + 1D CNN Layers:** These layers are responsible for extracting spatial and temporal features from each frame of the video. The 2D CNN layers capture spatial information within individual frames, while the 1D CNN layers capture temporal patterns across frames. This combination allows your model to learn both spatial and temporal representations simultaneously.
- **ReLU Activation Function:** Rectified Linear Unit (ReLU) activation functions are commonly used in neural networks to introduce non-linearity, helping the model learn complex patterns in the data.
- **Residual Networks (ResNets):** Residual networks are a type of deep neural network architecture that address the vanishing gradient problem by utilizing skip connections or shortcuts to jump over some layers. The residual blocks allow the model to learn residual mappings, which can be more easily optimized during training.

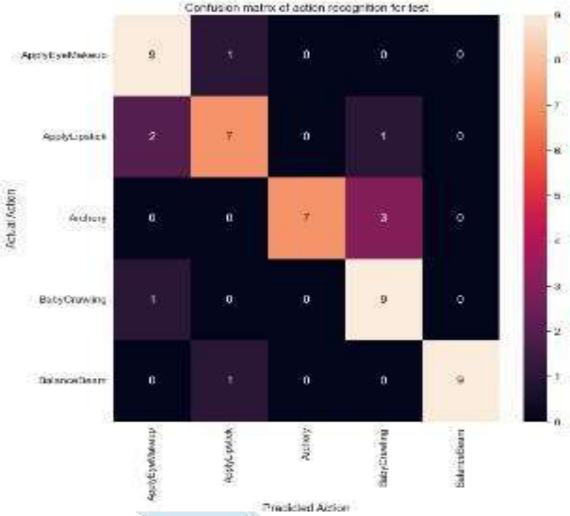
  You mentioned residual networks of varying depths (16, 32, 64, and 128 layers), indicating the depth of each residual network block. Deeper networks can capture more complex features but may also be more prone to overfitting if not regularized properly.
- **3D CNN Layers**: Unlike 2D CNNs, which operate on 2D spatial data (e.g., images), 3D CNNs operate on spatio-temporal data (e.g., videos). These layers are specifically designed to capture both spatial and temporal features simultaneously across multiple frames. They typically consist of 3D convolutional filters that slide across the spatial and temporal dimensions of the input data, extracting features from both space and time.
- **Average Pooling Layers**: Pooling layers are used to down sample the feature maps, reducing their spatial dimensions while retaining important information. Average pooling calculates the average value within each pooling region.
- **Flatten Layer:** This layer reshapes the output from the previous layer into a one-dimensional vector, which can then be fed into the subsequent dense (fully connected) layers.
- **Dense Layer:** Dense layers are fully connected layers where each neuron is connected to every neuron in the preceding layer. These layers are responsible for combining the extracted features and making predictions based on them.

A Custom 3D CNN model tailored for video labeling typically consists of multiple convolutional layers followed by pooling layers to extract relevant features. The architecture is customized based on the specific requirements of the video labeling task. The input to the CNN is a sequence of video frames, and the network learns spatial and temporal patterns through the convolutional and pooling operations. Architecture of CNN model shown in a below figure with each layer.

Proposed model has a multiple layers as above shown figure of 3D CNN model it takes 3D input data it convert 3 dimensional data into 2 dimensional data with necessary preprocessing methods model has pass through multiple layers of CNN model it gives **84% accuracy** with UCF101 dataset.

Input Layer: Accepts video data with shape (None, 10, HEIGHT, WIDTH, 3).

Conv2Plus1D Layer: Applies a 3D convolution followed by a 2D convolution with 16 filters and a kernel size of (3, 7, 7). Batch normalization and ReLU activation are applied afterward.



ResizeVideo Layer: Reduces the spatial dimensions of the video by half.

Residual Blocks:

Block 1: Uses a 3D convolution with 16 filters and a kernel size of (3, 3, 3).

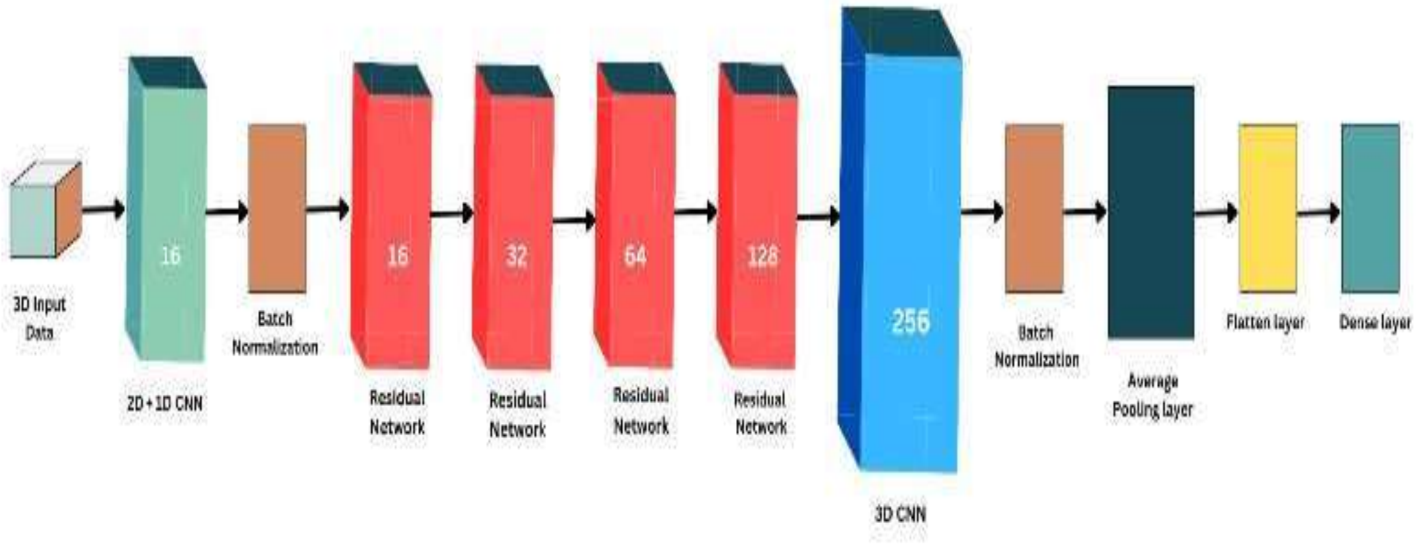
Block 2: Uses a 3D convolution with 32 filters and a kernel size of (3, 3, 3).

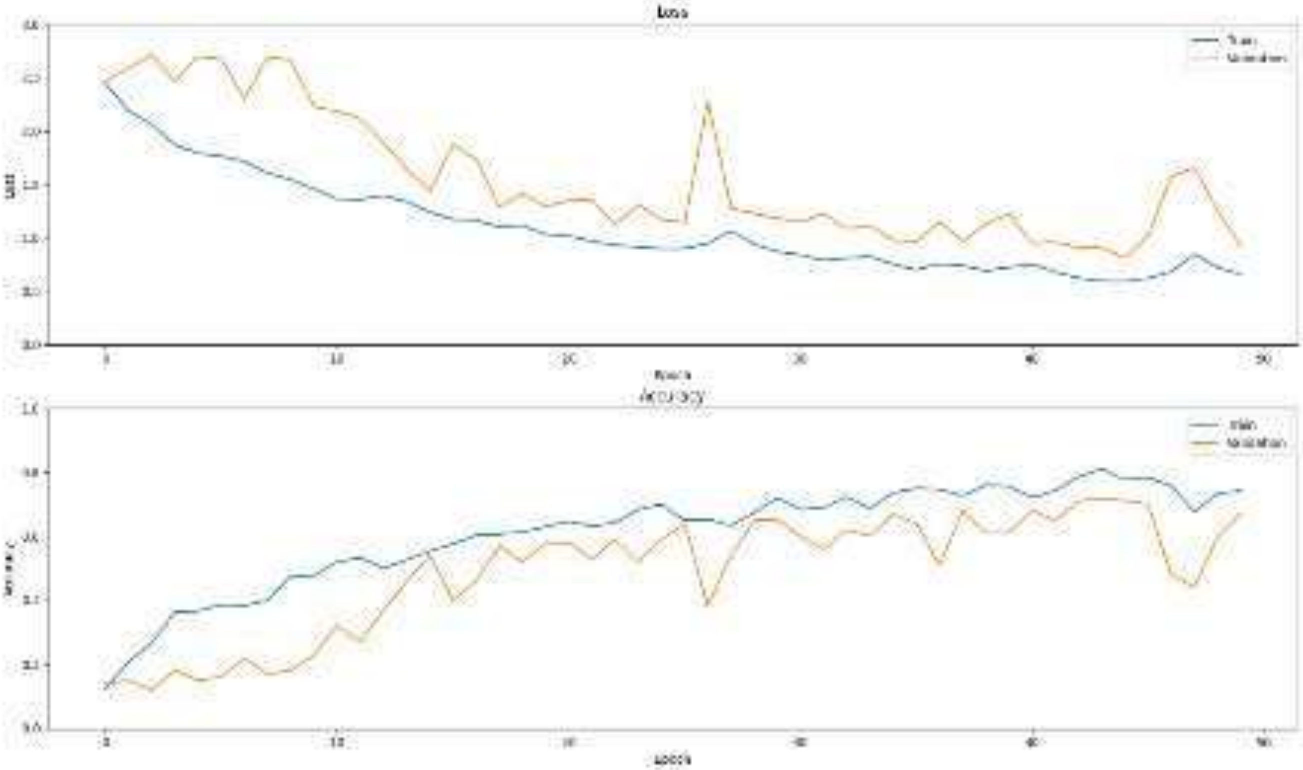
Block 3: Uses a 3D convolution with 64 filters and a kernel size of (3, 3, 3).

Block 4: Uses a 3D convolution with 128 filters and a kernel size of (3, 3, 3).

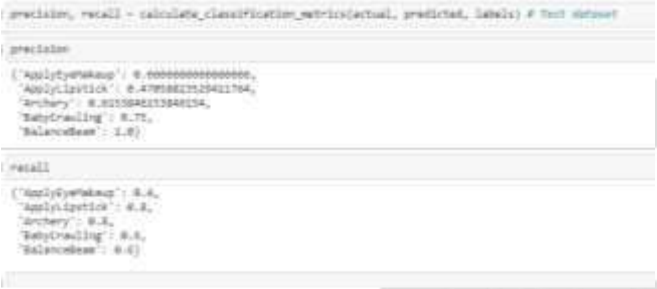
Conv3D Layer: Applies a 3D convolution with 256 filters and a kernel size of (3, 3, 3), followed by batch normalization and ReLU activation.

GlobalAveragePooling3D Layer: Performs global





average pooling across the spatial dimensions of the video.



```
precision, recall = calculate_classification_metrics(actual, predicted, labels) # Test dataset

precision

{'ApplyEyeMakeup': 0.6000000000000001,
 'ApplyLipstick': 0.47058823529411764,
 'Archery': 0.8153846153846154,
 'BabyCrawling': 0.75,
 'BalanceBeam': 1.0}

recall

{'ApplyEyeMakeup': 0.6,
 'ApplyLipstick': 0.8,
 'Archery': 0.8,
 'BabyCrawling': 0.6,
 'BalanceBeam': 0.6}
```

Flatten Layer: Flattens the output from the global average pooling layer.

Dense Layer: Produces the final output with 5 units (assuming it's a classification task).

Models	**Accuracy**	**Datasets**	**Number of Class es**
CNN	29%	UCF101	5
3D CNN	69%	UCF101	5
3D CNN with encoder decoder	67%	UCF101	5
CNN-RNN	67%	UCF101	5
Proposed Method (2+1)D CNN+ResNet+3D CNN model	**84%**	**UCF101**	**5**

- We have some other models comparison for our proposed model accuracy with same datasets and classes with different models of deep learning.

- The proposed 3D CNN model achieves the highest accuracy of 84% among the compared models.
- While the specific architectural details are not provided, it likely incorporates advancements such as deeper networks, improved regularization techniques, or novel architectural components tailored to the characteristics of the UCF101 dataset.
- The high accuracy demonstrates the effectiveness of the proposed model in capturing both spatial and temporal features in video data, leading to superior performance in action recognition tasks

• Conclusion

Proposed custom CNN model for identify best label for given input video .We have successfully developed and implemented advanced video labeling techniques for action recognition using the UCF-101 dataset. Our work integrates deep learning algorithms to accurately identify actions in videos and assign meaningful labels. The proposed custom CNN model significantly enhances the ability to determine the most appropriate label for a given input video.

This research combines established knowledge with innovative approaches, making a valuable contribution to the field. The outcomes of this work facilitate efficient video content analysis and enhance action recognition across diverse applications, ranging from sports analysis to entertainment.

Through research, my model achieved an accuracy of 84% on the UCF101 dataset using the first 5 classes as the raw data. This project has provided valuable insights into deep learning techniques for video classification, contributing to the field's advancement in action recognition and content analysis applications.

- **REFERENCES**

[1] *Host, K., & Ivašić-Kos, M. (2022). An overview of Human Action Recognition in sports based on Computer Vision. Heliyon.*

[2] *Kumar, M., Patel, A. K., Biswas, M., & Shitharth, S. (2023). Attention-based bidirectional-long short-term memory for abnormal human activity detection. Scientific Reports, 13(1), 14442.*

[3] *Kumar, V., Tripathi, V., & Pant, B. (2022). Learning unsupervised visual representations using 3d convolutional autoencoder with temporal contrastive modeling for video retrieval. International Journal of Mathematical, Engineering and Management Sciences, 7(2), 272-287.*

[4] *Ramesh, M., & Mahesh, K. (2022). Sports video classification framework using enhanced threshold based Keyframe Selection algorithm and customized CNN on UCF101 and Sports1-M dataset. Computational Intelligence and Neuroscience, 2022.*

[5] *Saoudi, E. M., & Jai-Andaloussi, S. (2021). A distributed Content-Based Video Retrieval system for large datasets. Journal of Big Data, 8(1), 1-26.*

[6] *Surek, G. A. S., Seman, L. O., Stefenon, S. F., Mariani, V. C., & Coelho, L. D. S. (2023). Video-Based Human Activity Recognition Using Deep Learning Approaches. Sensors, 23(14), 6384.*

[7] *Moniruzzaman, M., Yin, Z., He, Z., Qin, R., & Leu, M. C. (2021). Human action recognition by discriminative feature pooling and video segment attention model. IEEE Transactions on Multimedia, 24, 689-701.*

[8] *Video Representation Learning Using Discriminative Pooling.*

[9] *Qian, R., Meng, T., Gong, B., Yang, M. H., Wang, H., Belongie, S., & Cui, Y. (2021). Spatiotemporal contrastive video representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6964-6974).*

[10] *Pan, J., Chen, S., Shou, M. Z., Liu, Y., Shao, J., & Li, H. (2021). Actor-context-actor relation network for spatiotemporal action localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 464-474).*

[11] *Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey. International Journal of Computer Vision, 130(5), 1366-1401.*

[12] *Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., & Qiao, Y. (2023). Videomae*

v2: Scaling video masked autoencoders with dual masking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14549-14560).

[13] *Ramanathan, M., Yau, W. Y., & Teoh, E. K. (2014). Human action recognition with video data: research and evaluation challenges. IEEE Transactions on Human-Machine Systems, 44(5), 650-663.*

[14] *Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.*

[15] *Segalin, C., Perina, A., Cristani, M., & Vinciarelli, A. (2016). The pictures we like are our image: continuous mapping of favorite pictures into self-assessed and attributed personality traits. IEEE Transactions on Affective Computing, 8(2), 268-285.*

[16] *Hjelm, R. D., & Bachman, P. (2020). Representation learning with video deep infomax. arXiv preprint arXiv:2007.13278.*

[17] *Li, X., & Wang, L. (2023). ZeroI2V: Zero-Cost Adaptation of Pre-trained Transformers from Image to Video. arXiv preprint arXiv:2310.01324.*

[18] *Kumar, M., Patel, A. K., Biswas, M., & Shitharth, S. (2023). Attention-based bidirectional-long short-term memory for abnormal human activity detection. Scientific Reports, 13(1), 14442.*

[19] Payal, Parekh, and Mahesh M. Goyani. "A comprehensive study on face recognition: methods and challenges." *The Imaging Science Journal* 68, no. 2 (2020): 114-127.

[20] Awad, George, Keith Curtis, Asad Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado et al. "An overview on the evaluated video retrieval tasks at trecvid 2022." *arXiv preprint arXiv:2306.13118* (2023).

[21] Xiong, Bo, Haoqi Fan, Kristen Grauman, and Christoph Feichtenhofer. "Multiview pseudo-labeling for semi-supervised learning from video." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7209-7219. 2021.

[22] Lan, Zhenzhong, Yi Zhu, Alexander G. Hauptmann, and Shawn Newsam. "Deep local video feature for action recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 1-7. 2017.