IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

VISION VIGIL: A VIDEO CLASSIFIER

Manuraj Agarwal, Himanshu Pandey, Yashas Jaiswal, Tejas Srivastava
B.Tech Student
Department of Computer Science and Engineering
&
Department of Computer Science and Engineering (Data Science)
Inderprastha Engineering College, Sahibabad, Ghaziabad, India

Dr. Neeta Verma
Professor
Department of Computer Science and Engineering
Inderprastha Engineering College, Sahibabad, Ghaziabad, India

Ms. Garima Singh
Assistant Professor
Department of Computer Science and Engineering (Data Science)
Inderprastha Engineering College, Sahibabad, Ghaziabad, India

Abstract— Vision Vigil is a video classifier, leveraging state-of-the-art machine learning model and generative AI to tackle the challenges in categorizing digital multimedia. The project's primary objective is to efficiently distinguish between Safe for Work (SFW) and Not Safe for Work (NSFW) content, providing users with preemptive content warnings. By integrating a ML model and LLM model, Vision Vigil ensures accurate classification, offering a proactive approach to content analysis.

This innovative system goes beyond conventional video categorization, empowering diverse domains such as entertainment, security, education, and healthcare. Vision Vigil's adaptability enables seamless integration into various applications, offering a versatile solution for platforms prioritizing user safety. Through its sophisticated technology, Vision Vigil stands as a sentinel, cautioning viewers about the nature of video content and contributing to a more secure and informed digital landscape. In an era dominated by digital multimedia, Vision Vigil emerges as a crucial tool for enhancing content moderation and user experience.

Keywords— Video Classification, Audio Analysis, Text Classifier, Machine Learning, Multimedia Content, Deep Learning, Content Categorization, Generative AI, Gemini API Feature Extraction, Neural Networks, Sentiment Analysis, Image Recognition, Pattern Recognition, Content Moderation, Digital Media, Audio Segmentation, Video Analytics, Speech-to-Text, Image Classification, Language Models, Predictive Modeling, Information Retrieval, Topic Modeling, Audio Clustering, Document Classification, Image Segmentation, Supervised Learning, Unsupervised Learning, Multimedia Integration, Text Preprocessing, Video Labeling, Semantic Analysis, Textual Entailment, Video Content Warning, Multi-Modal Learning, Text Sentiment, Audio Pattern Recognition, Language Understanding, Video Summarization.

I. INTRODUCTION

Vision Vigil represents a groundbreaking leap in the realm of video classification, strategically crafted to navigate the intricate nuances of digital content categorization. With a primary focus on enhancing user experience and ensuring content appropriateness, this state-of-the-art project utilizes sophisticated machine learning techniques and neural networks and generative AI to categorize videos into two pivotal classifications: Not Safe for Work (NSFW) and Safe for Work (SFW). The core of Vision Vigil's functionality lies in its robust classifier, which processes videos by comprehensively analyzing images, audio, and text components. Vision Vigil employs machine learning classification techniques, harnessing neural networks to continually train and optimize the model. This dynamic adaptation ensures the system stays attuned to the evolving nature of digital content. The neural network architecture empowers the system to recognize patterns and subtle nuances, facilitating accurate classification into the NSFW or SFW categories. Beyond a binary classification, Vision Vigil generates detailed tags that can be seamlessly incorporated into the video description. This proactive feature serves as a cautionary measure, providing viewers with insights into the video's content before engagement. The inclusion of comprehensive tags not only enhances user awareness but also contributes to a more responsible and informed viewing experience.

II. LITERATURE SURVEY

Md Shofiqul and colleagues (2020) reviewed evaluates video classification methods, highlighting advantages, limitations, and emerging trends. It favours video-based approaches over text and audio, noting underutilization of text extraction. Balancing visual and audio feature extraction enhances classification. Audio-based solutions demand less computation. Innovative techniques involve segmenting images, setting thresholds, and employing diverse classification algorithms for movies, games, and event forecasting. Limitations include handling multiple features, deep learning's longer training time, and traditional machine learning's adaptability issues. Opportunities lie in classifying longer videos, recognizing multiple actions, establishing video correlations, categorizing multiple object actions, and exploring live streaming game video prediction as a burgeoning area in video classification research.

Andrej and colleagues (2014) The study focuses on large-scale video classification using Convolutional Neural Networks (CNNs). It emphasizes CNNs' ability to extract robust features from weakly-labelled data, outperforming feature-based methods consistently. While architectural details in time connectivity do not significantly affect performance, Slow Fusion models excel over early and late fusion alternatives. Surprisingly, even single-frame models display strong performance, hinting that local motion cues might not be crucial, challenging assumptions in dynamic datasets like Sports. Mixed-resolution architectures, combining low and high-resolution streams, enhance CNN speed without compromising accuracy. Transfer learning experiments demonstrate the generalizability of learned features. Future exploration aims for broader dataset categories, explicit treatment of camera motion, and investigating Recurrent Neural Networks for enhanced clip-to-video predictions.

Yinchong and colleagues (2017) integrated Tensor-Train Layers into Recurrent Neural Networks (RNNs) which revolutionizes their effectiveness in handling high-dimensional sequential data like videos. This enhancement significantly improves modelling performances compared to plain RNNs, offering simplicity and lightweight structures with far fewer parameters, enabling training and deployment on standard hardware and mobile devices. These tensorized models show promise in reducing the need for vast labelled data, typically expensive in the video domain. By enabling RNNs to directly process pixel-level video clips, this approach opens doors for applying successful RNN architectures from other domains like NLP to video modelling, including autoencoders, encoder-decoder networks for captioning, and attention-based models for improved classification. This breakthrough introduces RNNs as a viable solution for high-dimensional sequential data, bridging the gap where they previously struggled. The code for TT-RNN implementations and experiments is publicly accessible, fostering further exploration and development.

III. PROPOSED SYSTEM MODEL

The proposed methodology aims to comprehensively analyze video datasets by integrating visual, auditory, and textual features for robust classification. Initially, video datasets are categorized into SFW and NSFW. Audio transcript is then extracted from the video, capturing multiple dimensions of information and generating category tags using generative AI.

For feature extraction and image classification, the well-established Inception V3 model is utilized, leveraging the power of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) concepts and for audio classification transcripts are extracted and then with the help of generative AI model tags for video are generated.

Subsequently, the models are trained using the extracted features and labelled datasets, allowing them to learn and adapt to the specific characteristics of the input data. Once trained, the models collectively perform the classification of the video content, providing a holistic and multi-modal approach to understanding and categorizing diverse multimedia information. This comprehensive methodology facilitates a more nuanced and accurate analysis of video datasets across various dimensions.

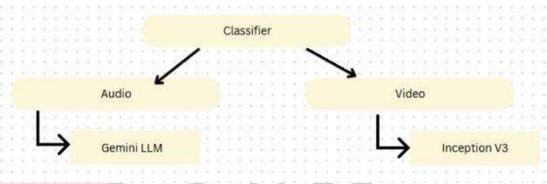


Figure 1 System Model

IV. METHODOLOGY

The proposed video classification methodology addresses the imperative need to effectively categorize the abundance of videos found in the real world. The primary objective is to classify videos into diverse categories such as athletics, films, amusing content, and educational material etc. To achieve this, two principal approaches are identified: audio-based and video-based. Additionally, a hybrid approach, integrating two or more methods, is suggested for a more nuanced and comprehensive classification strategy, depicting the Taxonomy of Video Classification Approaches.

The audio-centric methodology has gained notable attention due to its efficiency in time and energy conservation as compared to text-driven analyses. Central to this approach is the utilization of Speech Recognition technology, which facilitates the conversion of audio signals within video inputs into textual transcripts.

After the creation of these transcript files, they serve as prompts for Large Language Models (LLMs)(here Gemini), which have been extensively trained on vast datasets. Leveraging the contextual information encapsulated within the transcripts, these LLMs effectively categorize the content therein. This process not only expedites the analysis of audiovisual content but also enhances the accuracy.

Furthermore, the integration of Speech Recognition augments the accessibility and usability of audio-based data, allowing for seamless extraction and interpretation of auditory information. By harnessing the capabilities of LLMs, which encapsulate rich linguistic knowledge, the audio-based approach offers a robust framework for comprehensive content classification and analysis.

The video-based processing methodology comprises several sequential stages, commencing with the extraction of frames from the video source. Following this, the extracted frames undergo feature extraction, where pertinent visual characteristics are identified and distilled. Subsequently, a contextual stream is derived from these features, which is then forwarded to subsequent layers for classification.

Within these layers, Inception V3 architecture is employed, beginning with the input of fixed-size images (representative of video frames). These images are subjected to convolutional layers, which discern and encapsulate salient features from the input. Notably, the architecture facilitates the capture of features across diverse scales, enhancing its ability to comprehend nuanced visual information. Following convolution, pooling layers are employed to reduce the spatial dimensions of the feature maps, thereby condensing the information while retaining its critical aspects.

Upon pooling, the feature map undergoes flattening via fully connected layers, wherein the extracted features are organized into a one-dimensional array. Finally, the SoftMax layer operates on the output of the last fully connected layer, transforming it into a probability distribution, thereby facilitating effective classification.

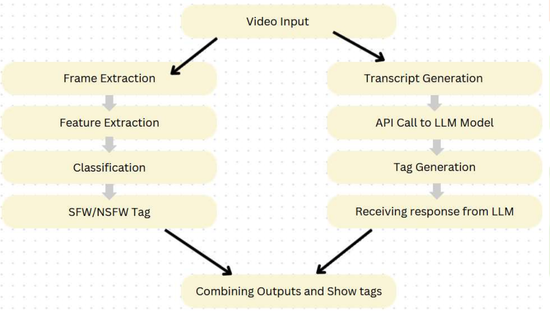
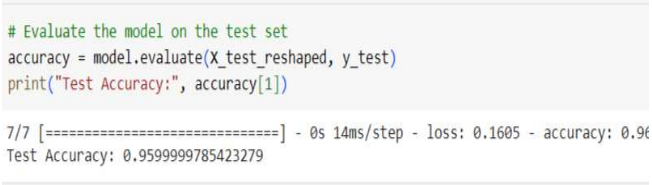


Figure 2 Proposed Methodology

V. TRAINING AND TESTING

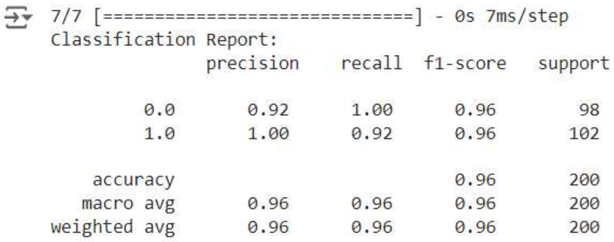
1) Train Accuracy:



```
# Evaluate the model on the test set
accuracy = model.evaluate(X_test_reshaped, y_test)
print("Test Accuracy:", accuracy[1])

7/7 [==============================] - 0s 14ms/step - loss: 0.1605 - accuracy: 0.96
Test Accuracy: 0.9599999785423279
```

Figure 3 Accuracy achieved after training

2) Classification Report:



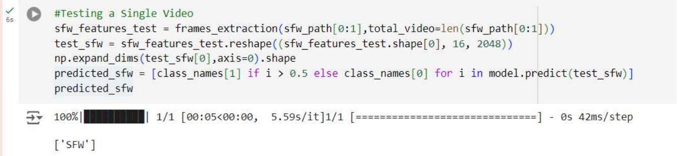
```
7/7 [==============================] - 0s 7ms/step
Classification Report:
              precision    recall  f1-score   support

         0.0       0.92      1.00      0.96        98
         1.0       1.00      0.92      0.96       102

    accuracy                           0.96       200
   macro avg       0.96      0.96      0.96       200
weighted avg       0.96      0.96      0.96       200
```

Figure 4 Various Scores achieved after training

3) Test Video:

Figure 5 Video given to test model

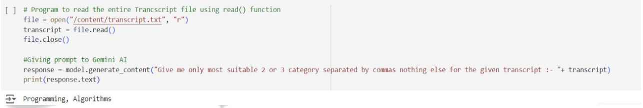
4) Video Analysis Output:



```
#Testing a Single Video
sfw_features_test = frames_extraction(sfw_path[0:1],total_video=len(sfw_path[0:1]))
test_sfw = sfw_features_test.reshape((sfw_features_test.shape[0], 16, 2048))
np.expand_dims(test_sfw[0],axis=0).shape
predicted_sfw = [class_names[1] if i > 0.5 else class_names[0] for i in model.predict(test_sfw)]
predicted_sfw

100%|██████████| 1/1 [00:05<00:00,  5.59s/it]1/1 [==============================] - 0s 42ms/step

['SFW']
```

Figure 6 Video Analysis Result

5) Audio Analysis Output:



```
# Program to read the entire Trancscript file using read() function
file = open("/content/transcript.txt", "r")
transcript = file.read()
file.close()

#Giving prompt to Gemini AI
response = model.generate_content("Give me only most suitable 2 or 3 category separated by commas nothing else for the given transcript :- "+ transcript)
print(response.text)

Programming, Algorithms
```

Figure 7 Audio Analysis Result

VI. CHALLENGES AND LIMITATIONS

1) **Computational Demands**: Analysing video frames requires significant computational resources, especially for high-resolution videos or real-time processing, leading to slower inference times.

2) **Feature Extraction and Representation**: Selecting relevant features and effectively representing them for classification is crucial. Extracting meaningful features from raw video data while avoiding information loss is a challenge.

3) **Class Imbalance and Annotation Errors**: Imbalanced datasets or errors in annotations can impact model performance, leading to biases and inaccuracies in classification.

4) **Variability in Audio Content:** Audio data can vary significantly in terms of background noise, pitch, tempo, and accent, making it challenging to develop models robust to these variations.

5) **Lack of Labelled Data:** Acquiring labelled audio datasets for training can be limited, especially for specialized domains or less common audio categories, hindering the development of accurate classifiers.

6) **Dimensionality:** Representing audio data effectively by extracting relevant features while managing high-dimensional data poses a challenge. Selecting suitable features for differentiating classes is crucial.

7) **Ambiguity and Contextual Understanding:** Ambiguous language, nuances, sarcasm, and varying contexts in text make accurate classification challenging, as models might struggle to comprehend subtleties in meaning.

8) **Multilingual and Cross-Lingual Challenges:** Developing classifiers that can handle multiple languages or translate text for cross-lingual applications presents significant challenges in maintaining accuracy and consistency.

9) **Domain Adaptation:** Models trained on one domain might not generalize well to different domains due to differences in language use, vocabulary, or writing style.

VII. PROPOSED ENHANCEMENTS

1) **Temporal Modeling:** Improved architectures for capturing temporal relationships across frames, such as attention mechanisms or spatiotemporal convolutions, enhance video understanding.
2) **Multi-Modal Fusion:** Integrating multiple modalities (text) for a richer understanding of video content improves classification accuracy, especially in complex scenarios.
3) **Transfer Learning and Pre-Trained Models:** Fine-tuning models pre-trained on massive text corpora for specific tasks improves classification accuracy, especially with limited labelled data.
4) **Attention Mechanisms:** Integrating attention mechanisms to focus on important segments of text for better feature representation and understanding.
5) **Improved Feature Representation:** Exploring novel methods for feature extraction, such as Mel-frequency cepstral coefficients (MFCCs), spectrograms, or wavelet transforms, to capture unique audio characteristics.
6) **Deep Learning Architectures:** Employing deep learning architectures like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) for learning temporal dependencies and hierarchical representations.
7) **Transfer Learning:** Transferring knowledge from models trained on large audio datasets like Audio Set or ESC-50 to enhance performance on specific audio classification tasks.
8) **Robustness to Noise:** Developing models resilient to background noise or environmental variations to ensure accurate classification in diverse audio environments.

VIII. CONCLUSION

In conclusion, the integration of advanced technologies such as the Inception V3 model for video classification, speech-to-text conversion for audio analysis, and generative AI models marks a significant advancement in the realm of multimedia content processing. By leveraging state-of-the-art machine learning algorithms, our video classifier not only accurately categorizes visual content but also enhances user interaction through the extraction of speech and subsequent generation of relevant tags via generative AI.

The utilization of the Inception V3 model allows for robust video classification, enabling the system to effectively categorize diverse visual content with high accuracy. Additionally, by extracting speech from the audio component of the multimedia content and converting it into a transcript file, our system bridges the gap between auditory and textual data, further enriching the analysis process.

The incorporation of generative AI in response to the transcript file not only adds a layer of semantic understanding but also provides personalized and contextually relevant tags, enhancing the overall user experience. This adaptive approach ensures that the generated tags are tailored to the specific content of the video, thereby improving searchability and discoverability.

Overall, our research showcases the potential of combining cutting-edge machine learning techniques to create a comprehensive multimedia analysis system. As technology continues to evolve, further advancements in video classification, audio analysis, and generative AI promise to revolutionize the way we interact with and interpret multimedia content.

IX. REFERENCE

- D. D. Lewis, "A Sequential Algorithm for Training Text Classifiers: Corrigendum and Additional Data," AT&T Bell Laboratories, Murray Hill, NJ 07974, USA, 1995.
- J. Jiang, Z. Li, J. Xiong, R. Quan, Q. Lu, and W. Liu, "Tencent AVS: A Holistic Ads Video Dataset for Multi-Modal Scene Segmentation," Tencent Data Platform, Shenzhen 518057, China, 2022.
- H. Shen, S. Han, M. Philipose, and A. Krishnamurthy, "Fast Video Classification via Adaptive Cascading of Deep Models," University of Washington, Rubrik, Inc., Microsoft Research, 2017.
- S. Pentyala, R. Dowsley, and M. De Cock, "Privacy-Preserving Video Classification with Convolutional Neural Networks," 2021.
- A. u. Rehman, S. B. Belhaouari, M. A. Kabir, and A. Khan, "On the Use of Deep Learning for Video Classification," Artificial Intelligence and Intelligent Systems Research Group, School of Innovation, Design and Technology, Mälardalen University, Högskoleplan 1, 722 20 Västerås, Sweden, 2023.

- D. D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka, "Training algorithms for linear text classifiers," in Proceedings of the 1996 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 96), Zurich, Switzerland, 1996.
- L. Jing, T. Parag, Z. Wu, Y. Tian, and H. Wang, "VideoSSL: Semi-Supervised Learning for Video Classification," The City University of New York, Comcast Applied AI Research, 2021.
- W. Zhou, A. Vellaikal, and C.-C. J. Kuo, "Rule-based Video Classification System for Basketball Video Indexing," in Proceedings of [Conference Name], 2002.
- J. Wang, Q. Wu, H. Deng, and Q. Yan, "Real-Time Speech/Music Classification with a Hierarchical Oblique Decision Tree," in Proceedings of [Conference Name], 2008.
- Urbano Rome, "Emotion Recognition Based on the Speech Using a Naive Bayes Classifier," 2016.
- S. Zha, F. Luisier, W. Andrews, "Exploiting Image-trained CNN Architectures for Unconstrained Video Classification," Northwestern University, Evanston, IL, USA, and Raytheon BBN Technologies, Cambridge, MA, USA, 2015.
- Y. Xu, "A Sports Training Video Classification Model Based on Deep Learning," 2021.
- N. Casagrande, D. Eck, and B. Kégel, "Geometry in Sound: A Speech/Music Audio Classifier Inspired by an Image Classifier," 2005.
- S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A Large-Scale Video Classification Benchmark," 2016.
- Y. Yang, D. Krompass, and V. Tresp, "Tensor-Train Recurrent Neural Networks for Video Classification," 2017.
- A review on Video Classification with Methods, Findings, Performance, Challenges, Limitations and Future Work Md Shofiqul Islam1,2 , Shanjida Sultana3 , Uttam kumar Roy4 , Jubayer Al Mahmud5,2020
- M. S. Islam, S. Sultana, U. K. Roy, J. A. Mahmud, "A Review on Video Classification with Methods, Findings, Performance, Challenges, Limitations, and Future Work," 2020.
- S. Bhardwaj, M. Srinivasan, M. M. Khapra, "Efficient Video Classification Using Fewer Frames," 2019.
- Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," 2014.

IJCRT