IPEC Journal of Science & Technology, Vol. 02 (02), Dec 2023
ISSN: 2583-3286(Online)

Vision vigil: A Video classifier

Dr. Sandhya Umrao*, Dr. Neeta Verma, Manuraj Agarwal**, Himanshu Pandey**, Yashas Jaiswal**, Tejas Srivastava****

**Department. of Computer Science and Engineering, Noida Institute of Engineering & Technology, Greater Noida*

***Department. of Computer Science and Engineering, Inderprastha Engineering College, U.P, India*

***Abstract*: Vision Vigil stands at the forefront of video classification, leveraging state-of-the-art machine learning algorithms to tackle the nuanced 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 cutting-edge ML 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.**

***Keywords*: Video Classification, Audio Analysis, Text Classifier, Machine Learning, Multimedia Content, Deep Learning**

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 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 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 favors 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-labeled data, outperforming feature-based methods consistently. While architectural details in time connectivity don't 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.

Date of Submission: 29 Nov 2023
Date of Acceptance: 20 Dec 2023
Corresponding Author: Manuraj Agarwal
(e-mail: manu.ag002@gmail.com

This enhancement significantly improves modeling 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 labeled 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 modeling, 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 various classes, providing a diverse set of content for analysis. Frames, audio, and text are then extracted from the videos, capturing multiple dimensions of information. To enhance the analysis, two distinct data streams, termed fovea and context, are employed. The fovea stream focuses on specific regions of interest within frames, while the context stream considers the broader visual context.

For feature extraction and image classification, the well-established VGG16 model is utilized, leveraging the power of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) concepts. The VGGish model is employed for audio classification, specializing in extracting relevant features from the audio data. Simultaneously, the XLNet model is applied for text classification, emphasizing the textual content within the

videos. Subsequently, the models are trained using the extracted features and labeled 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.

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. To achieve this, three principal approaches are identified: textbased, 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.

Figure 1: Proposed model

In the text-based approach, the generation and evaluation of video texts play a pivotal role. This involves extracting visible text or converting speech into text. Computer derived text encompasses elements like playing board details, jersey numbers, and on-screen subtitles, which are extracted using Optical Character Recognition (OCR). Similarly, voice recognition is employed for extracting text from speech, commonly used for subtitles and closed captions, accommodating various audio forms such as pet sounds or songs. The audio-based approach gains prominence due to its efficiency in terms of time and energy consumption compared to text-based analysis. Audio signals are sampled, and relevant characteristics are extracted for analysis, utilizing both the time and frequency domains. This approach not only requires less storage space but also proves to be computationally efficient in processing audio features.

The video-based approach relies on human interpretation of visual information, capturing the inherent complexity of videos. Visual characteristics are extracted from image sequences or video files, focusing on elements such as color, motion, shot time, and other visual features. Recognizing a video as a collection of frames, this approach considers lighting, movement, background, and video speed details for classification. A comparative analysis among the video classification approaches,

ISSN: 2583-3286(Online)

offering a detailed examination of the strengths and limitations of each method. This comparative insight assists in making informed decisions regarding the suitability of a particular approach based on the application's requirements. The comprehensive nature of this methodology ensures that it can effectively handle the intricacies of video content across various dimensions, offering a robust and adaptable solution to video classification challenges.

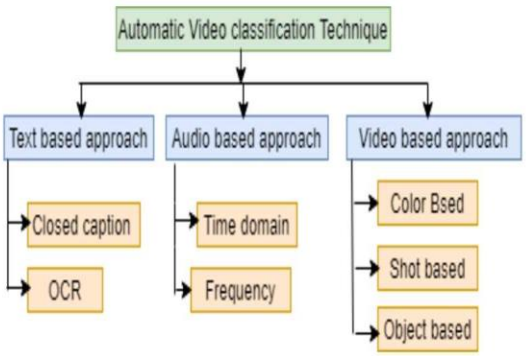

Figure 2: Approaches of classifications

V. CHALLENGES AND LIMITATIONS

1) Computational Demands: Analyzing 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 Labeled Data: Acquiring labeled audio datasets for training can be limited, especially for specialized domains or less common audio categories, hindering the development of accurate classifiers.
6) Dimensionality and Feature Extraction: 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) Evaluation Metrics: Defining appropriate evaluation metrics that accurately measure the performance of audio classifiers across different tasks and domains can be complex.
8) 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.
9) 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.
10) Domain Adaptation: Models trained on one domain might not generalize well to different domains due to differences in language use, vocabulary, or writing style.

VI. PROPOSED ENHANCEMENTS AND ADAPTATIONS

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 (audio, text, visual) for a richer understanding of video content improves classification accuracy, especially in complex scenarios.
3) Transfer Learning and Pre-Trained Models: Finetuning models pre-trained on massive text corpora for specific tasks improves classification accuracy, especially with limited labeled data.
4) Contextual Embeddings: Utilizing transformer-based models like BERT, GPT, or RoBERT a for contextual understanding and capturing nuanced semantics within text data.
5)Attention Mechanisms: Integrating attention mechanisms to focus on important segments of text for better feature representation and understanding.
6) 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.
7) Deep Learning Architectures: Employing deep learning architectures like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) for learning temporal dependencies and hierarchical representations.
8) 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.
9) Robustness to Noise: Developing models resilient to background noise or environmental variations to ensure accurate classification in diverse audio environments.

VII. CONCLUSION

In conclusion, this comprehensive review assesses video classification methodologies, shedding light on their strengths, limitations, and emerging trends. Video-based approaches are favored, emphasizing the underutilization of text extraction and the enhanced classification achieved by balancing visual and audio feature extraction.

The study identifies innovative techniques, such as image segmentation and diverse classification algorithms, showcasing their applicability in domains like movies, games, and event forecasting. Despite challenges in handling multiple features and adaptability issues in traditional machine learning, the opportunities in classifying longer videos, recognizing multiple actions, and exploring live streaming game video prediction are promising avenues for future research. The focus on large-scale video classification using Convolutional Neural Networks (CNNs) demonstrates their superiority in extracting robust features from weakly labeled data. The study reveals the efficiency of Slow Fusion models and mixed-resolution architectures in enhancing CNN speed without compromising accuracy. Transfer learning experiments underscore the generalizability of learned features, opening avenues for broader dataset categories. The integration of Tensor-Train Layers into Recurrent Neural Networks (RNNs) marks a breakthrough, revolutionizing their effectiveness in handling high-dimensional sequential data like videos. This approach introduces RNNs as a viable solution for video modeling, offering simplicity and lightweight structures with broader applicability. The publicly accessible code for TT-RNN implementations encourages further exploration and development, hinting at a transformative future in video classification research.

REFERENCE

[1].D. D. Lewis, "A Sequential Algorithm for Training Text Classifiers: Corrigendum and Additional Data,"AT&T Bell Laboratories, Murray Hill, NJ 07974,USA, 1995.

[2].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.

[3].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.

[4] S. Pentyala, R. Dowsley, and M. De Cock, "Privacy Preserving Video Classification with Convolutional Neural Networks," 2021.

[5]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.

[6] 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.

[7] 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.

[8] W. Zhou, A. Vellaikal, and C.-C. J. Kuo, "Rule-based Video Classification System for Basketball Video Indexing," in Proceedings of [Conference Name], 2002.

[9] 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

[10] Urbano Rome, "Emotion Recognition Based on the Speech Using a Naive Bayes Classifier," 2016.

[11]S. Zha, F. Luisier, W. Andrews, "Exploiting Imagetrained CNN Architectures for Unconstrained Video Classification," Northwestern University, Evanston, IL, USA, and Raytheon BBN Technologies, Cambridge, MA, USA, 2015.

[12]Y. Xu, "A Sports Training Video Classification Model Based on Deep Learning," 2021.

[13] N. Casagrande, D. Eck, and B. Kégel, "Geometry in Sound: A Speech/Music Audio Classifier Inspired by an Image Classifier," 2005.

[14]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.

[15]Y. Yang, D. Krompass, and V. Tresp, "Tensor-Train Recurrent Neural Networks for Video Classification," 2017.

[16] 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

[17]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.

[18] S. Bhardwaj, M. Srinivasan, M. M. Khapra, "Efficient Video Classification Using Fewer Frames," 2019.

[19 Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," 2014.