Assessing TikTok Videos Content of Tobacco Usage by Leveraging Deep Learning Methods

Naga VS Raviteja Chappa[1], Charlotte McCormick[2], Susana Rodriguez Gongora[2], Page Daniel Dobbs[2], Khoa Luu[1]
[1]Dept. of EECS, University of Arkansas
[2]Center for Public Health and Technology, University of Arkansas
{nchappa, cem044, sr069, pdobbs, khoaluu}@uark.edu

***Abstract*—This study addresses concerns surrounding the inadvertent promotion of tobacco-related products on TikTok by introducing an efficient deep learning-based video analysis system. Our approach focuses on categorizing TikTok videos based on tobacco-related cues, including content related to e-cigarettes, vapes, cigarettes, various tobacco flavors, and accessories that may bypass tobacco control policies.**

The proposed two-stage process begins with the extraction of essential cues using speech-to-text, Optical Character Recognition (OCR), and video classification techniques. This initial phase comprehensively captures textual and visual information associated with tobacco products, forming the foundation for understanding video content. Subsequently, in the second stage, the extracted cues are integrated into a vision-language model alongside the input video. This stage trains the model to analyze contextual nuances, achieving a detailed understanding of the nuanced elements associated with tobacco promotion on TikTok.

The system classifies input videos into predefined classes (cigarette, e-cigarette/vapes, pouches, or others) and provides detailed analyses. This capability enables a granular examination of diverse tobacco-related content on TikTok, proving valuable for regulatory agencies like the FDA in quickly identifying potential illegal promotion and sales of non-compliant tobacco products and accessories.

I. Introduction

In the evolving landscape of digital content, the surge in the popularity of short-form videos on platforms like TikTok has ushered in a new era of user-generated content. As video-sharing platforms continue to redefine how information is consumed, there is an increasing demand for advanced deep learning applications that can comprehend the intricacies of video content.

While language models have demonstrated remarkable capabilities in processing and understanding textual data, their adaptation to video understanding remains a nascent field. Current state-of-the-art Large Language Model (LLM) architectures [1], [2], [3], [4], [5] excel in natural language processing tasks but are inherently limited by their lack of explicit support for video inputs. This limitation becomes particularly pronounced in the context of platforms like TikTok, where video content takes center stage in communication and expression.

TikTok, with its vast and diverse user base, has become a microcosm of content ranging from entertainment to social commentary. However, the inadvertent promotion of tobacco-related products on this platform has raised significant concerns. Existing LLMs are ill-equipped to tackle this issue as they are designed to operate primarily on text, neglecting the rich visual information embedded in video content.

In response to this gap, our proposed approach ventures into the intersection of deep learning and video analysis. Focused on the nuanced domain of tobacco-related content, we present a two-stage end-to-end system that not only extracts essential cues from TikTok videos but also trains a vision-language model to comprehend and analyze the contextual intricacies within the video content. By combining the power of deep learning with the unique challenges presented by TikTok's video-centric format, our approach aims to provide a comprehensive solution to the complex issue of tobacco promotion in this dynamic digital space.

This paper unfolds the proposed methodology, and expected results after the implementation, shedding light on the imperative role that advanced deep learning applications can play in addressing emerging challenges in the realm of video content moderation.

II. Related Work

Recent advancements in video analysis [6], [7] have predominantly employed traditional deep learning algorithms, but their efficacy is limited by the absence of language comprehension. Recognizing this gap, there is a shift towards integrating text modalities [8] to enhance understanding capabilities. This underscores the need to evolve existing algorithms, advocating for joint training with Large Language Models (LLMs) [1] to achieve superior performance and nuanced comprehension. Our work builds upon this paradigm shift, proposing a novel approach that seamlessly integrates video and language understanding. Through leveraging the power of LLMs, our framework bridges the gap between visual and textual cues, advancing the state-of-the-art in video analysis. This convergence enables a holistic interpretation of video content, opening new avenues for more accurate, context-aware applications.

III. Proposed Methodology

Our innovative two-stage methodology, illustrated in Fig. 1, unfolds as a sophisticated, end-to-end process meticulously designed for the comprehensive analysis and categorization of TikTok videos centered around tobacco-related cues. This

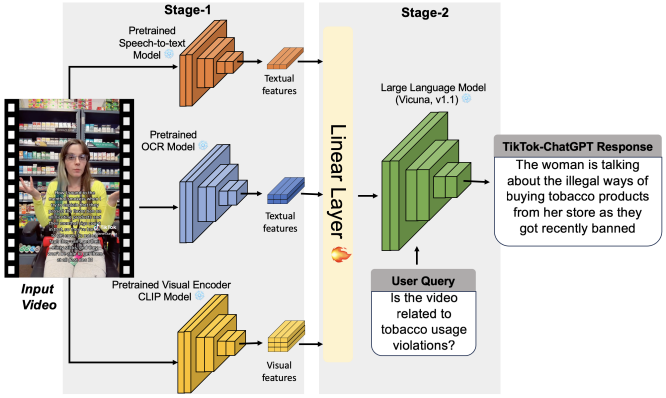

Fig. 1. **Framework of TikTok-VideoChatGPT**. Our framework utilizes the CLIP visual encoder to extract spatiotemporal features from TikTok videos. These features are refined through a learnable linear layer and projected into Large Language Models (LLMs) using the 7B-parameter Vicuna-v1.1 model. Initialization is performed with weights from LLaVA [1], ensuring a robust foundation for subsequent video analysis stages.

holistic approach seamlessly integrates both textual and visual components inherent in video content, providing a robust framework for deeper understanding.

Stage 1 - Feature Extraction: In the inaugural stage, we employ a multi-faceted strategy to extract vital cues from TikTok videos, ensuring a holistic perspective. Utilizing cutting-edge pretrained techniques, including speech-to-text [9], Optical Character Recognition (OCR) [10], and a powerful video encoder [11], our objective is to capture spoken and written features related to tobacco in their entirety along with the video features. This comprehensive feature extraction ensures a profound understanding of textual and visual information associated with tobacco products, establishing the groundwork for subsequent in-depth analysis.

Stage 2 - Vision-Language Integration: Moving to the second stage, we seamlessly integrate the extracted textual and visual cues into a vision-language model, enhancing our model's contextual comprehension. Drawing inspiration from the success of models like Video-ChatGPT [12], our approach aims to train the model to adeptly comprehend the nuanced aspects specific to tobacco-related content on TikTok. This integration empowers our model to conduct a nuanced and detailed analysis by harmonizing the strengths of language models with rich visual information.

Through the simultaneous interpretation and analysis of both textual and visual features, facilitated by a learnable linear layer, our model aspires to achieve an elevated understanding of the nuanced elements associated with tobacco promotion on TikTok. The result is a versatile system poised to address the complex challenges posed by diverse forms of tobacco-related content on this dynamic social media platform.

IV. Conclusion

In conclusion, our work presents a pioneering methodology that amalgamates video analysis with advanced language understanding models, exemplified by CLIP. By unifying spatial and temporal features extracted from videos with textual cues, our approach achieves a comprehensive interpretation of content. This symbiotic relationship between visual and language modalities elevates the potential for nuanced comprehension and robust performance in video analysis tasks. As we move forward, this integration lays the foundation for more sophisticated and contextually aware applications, addressing the evolving demands of multimedia content understanding in the digital age. Our study contributes to the ongoing dialogue on the synergy between deep learning and video analysis, underscoring the significance of interdisciplinary approaches in advancing the frontiers of artificial intelligence.

Significance and Potential Impact: This methodology, by virtue of its fusion of deep learning and video analysis, holds substantial promise in addressing the challenges posed by tobacco-related content on TikTok. The deep analysis enabled by our approach is not only beneficial for content creators and platform moderators but also crucial for regulatory bodies such as the FDA. Rapid identification and classification of videos promoting tobacco products can aid in enforcing compliance with federal, state, and local regulations, ensuring a safer and more responsible digital environment.

References

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[2] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[3] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[4] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[5] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[6] N. V. Chappa, P. Nguyen, A. H. Nelson, H.-S. Seo, X. Li, P. D. Dobbs, and K. Luu, "Spartan: Self-supervised spatiotemporal transformers approach to group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5157–5167.

[7] ——, "Sogar: Self-supervised spatiotemporal attention-based social group activity recognition," *arXiv preprint arXiv:2305.06310*, 2023.

[8] N. V. Chappa, P. Nguyen, P. D. Dobbs, and K. Luu, "React: Recognize every action everywhere all at once," *arXiv preprint arXiv:2312.00188*, 2023.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[10] M. Ye, J. Zhang, S. Zhao, J. Liu, T. Liu, B. Du, and D. Tao, "Deepsolo: Let transformer decoder with explicit points solo for text spotting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 348–19 357.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[12] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv preprint arXiv:2306.05424*, 2023.