Graph Attention Network (GAT) to Understand Short Video Hook Effectiveness

Arian Kalantari
48549819

Macquarie University

Abstract: In the age of short-form video platforms like Instagram Reels and TikTok, marketers face increasing pressure to capture viewer attention within the first few seconds. These brief moments, commonly referred to as "hooks," are critical for engagement but difficult to evaluate before publication. This paper addresses the challenge of pre-publication hook evaluation by proposing a multimodal Graph Neural Network (GNN) approach. The model segments videos into fine-grained temporal slices and extracts multimodal features—including visual, audio, text, camera movement, and actor motion—from each segment. A graph is constructed to capture both intra-video structure and inter-video relationships, such as shared trends or similar engagement metrics. A Graph Attention Network (GAT) is applied to this structure to identify and score hook effectiveness based on patterns learned from high-performing content. The proposed approach offers a scalable and interpretable way to help marketers validate content before release. Beyond marketing, these insights may also inform applications in education and cognitive science, contributing to a broader understanding of attention dynamics in digital media.

Keywords: short-form video, hook prediction, multimodal graph neural network, video engagement, content evaluation.

1 Introduction

1.1 Background

The digital marketing industry has grown significantly, fueled by the rise of social network apps and active participation of both users and business owners in content creation. [1] One of the most influential formats contributing to the growth is the short-form video, such as Instagram Reels or TikTok. [2] These videos align with the modern audiences' fast-paced consumption habits, urging marketers to capture viewers' attention within the first few seconds. [3]

This nature has led to increased awareness of the term "hook" in short-form video content creation strategy. The concept of "hooks" in video content refers to those initial elements designed to immediately capture the viewer's attention and pique their inter-

est, compelling them to continue watching. Hooks can take many forms—visual, auditory or textual—may include scenes, dynamic editing, movements, impactful narrations and on-screen text. [4] The concept of "hooks" in video content refers to those initial elements designed to immediately capture the viewer's attention and pique their interest, compelling them to continue watching. Without effective hooks, content often fails to gain meaningful user engagement.

Research suggests that the effective hooks are crucial for maximising engagement, especially in the context of short-form videos where viewers have limited attention spans and numerous content options. [5] These hooks play a vital role in narrative transportation and viewer engagement, drawing the audience into the video's content and making them more receptive to the intended message. [6] As such, views, likes, and comments are commonly used as post-hoc measure to assess the success of content. However, predicting engagement based on hook quality remains an unresolved challenge in digital content creation.

1.2 Motivation

In today's oversaturated digital media landscape, capturing and retaining viewer attention has become a major challenge for marketers and content creators. [2] While short-form videos have proven to be a powerful medium for engagement, only a small fraction of such content achieves meaningful interaction. A key reason is the weakness of the "hook" within 3-4 second window that determines whether viewers continue watching or scroll past. [5][6]

While various analytics tools—Instagram Reels Insights exist for evaluating the performance or suggesting trending patterns, most of them operate post-publication, offering insights based on engagement metrics after the content is released. Other tools—TikTok Creator Centre, Wisecut—are available for content creators pre-publication, however, they operate based on trends in million videos to suggest video clips. This paper addresses the need for a pre-publication evaluation framework by proposing a Graph Neural Network (GNN)-based approach that analyses multimodal signals from short-form videos and learns from existing engagement patterns to predict hook effectiveness. GNNs offer the ability to model relationships across videos, uncover common engagement patterns, and provide creators with actionable feedback before content is released.

Beyond improving content performance, the ability for AI to predict attention dynamics raises ethical questions. If used responsibly, such models could enhance educational media and promote constructive engagement. However, they also pose risks of manipulative optimisation. Therefore, this work aims to develop a scalable and ethically aware method for evaluating hook effectiveness in short-form videos.

2 Related Work

Numerous studies have explored the application of AI in video content analysis, particularly focusing on engagement prediction, multimodal analysis, and graph-based modeling. This section outlines three methods from the literature that are relevant to hook detection in short-form videos.

2.1 Multimodal Emotion Recognition using Deep Learning

A common approach in analysing video content is using multimodal deep learning to recognise emotional cues from video, audio, and text. Hooks that trigger emotional reactions—such as surprise, curiosity, or humor—should be analysed to better understand their impact on viewer engagement. For example, Tzirakis et al. (2017) introduced the Multimodal Emotion Recognition using Deep Neural Networks for combining modalities to improve emotion detection. Their model successfully integrated visual expressions, voice tone, and subtitles to estimate viewer reaction. [7]

However, such models treat each video as an independent instance, ignoring cross-video relational structures—such as market trends (e.g., dance or transition styles), user expectations (e.g., recommendation algorithms), and platform dynamics—that can significantly influence content performance.

2.2 Vision Transformer-Based Video Understanding

Transformer-models such as VideoBERT (Sun et al., 2019) use transformers to learn joint representations of video frames and transcripts, which is particularly useful for understanding short, structured segments (e.g. hooks). These models, by applying multimodality in learning, performed better than single modal-based models in generating semantic embeddings that capture narrative structure. [8] However, these models often require large-scale labeled datasets and treat each video as an independent instance. This limits their ability to model relational patterns across videos, such as trending formats or repeated hook structures, and they often lack integration with engagement-specific metadata.

2.3 Graph-Based User Engagement Prediction

Graph Neural Networks (GNNs) have emerged as powerful tools for modeling complex relationships in social media data. [9] Models like SocialGCN (Wu et al., 2019) leverage GNNs to learn from user-content interaction graphs, effectively capturing relational information and generalising across sparse data.

However, above mentioned work has predominantly focused on recommendation systems, often overlooking the content-level analysis of hooks in short-form videos. SocialGCN only uses user-item interaction graphs, ignoring content-specific modality information. Additionally, many models neglect the fusion of multimodal inputs—such as images, audio, and text—within node representations. [10]

Recent surveys highlight the necessity of integrating multimodal data and temporal dynamics into GNN frameworks to enhance user behavior prediction (Singh, 2024).

These insights suggest an opportunity to explore multimodal graph neural networks that can model both cross-video and inter-video relationships, potentially yielding more robust insights into patterns of viewer engagement.

3 Identified Methodologies

3.1 Problem Statement

Despite the growing dominance of short-form videos in digital marketing, reliable pre-publication evaluating tools for the effectiveness of video hooks and score prediction is not widely explored.

The central problem this study addresses is the absence of a content-level, predictive framework that can assess hook effectiveness using multimodal and temporal signals. Particularly, effective hooks often share latent patterns across camera movement, actor dynamics, transitions, speech, and on-screen text—patterns which are overlooked by existing models that treat videos in isolation or rely solely on static features.

This research paper proposes learning these patterns by segmenting videos into temporal slices and modeling them using graph-based structures. Each slice is represented by a rich set of multimodal features, while relationships between slices across different videos are encoded as edges in a graph. The goal is to train a model that can learn from previously successful content and predict the engagement potential of new video hooks before they are published, enabling marketers to make data-driven creative decisions.

3.2 Proposed Methods

To solve the stated problem, we propose a Graph Attention Network (GAT) that combines content-level feature learning with relational modeling across short-form videos. The process begins by segmenting each input video into short temporal slices, typically covering the initial 0–3 seconds or smaller intervals, to capture the critical hook region. This segmentation maintains temporal granularity, which is essential for detecting rapid visual or narrative shifts that influence viewer retention.

From each segment, multimodal features are extracted using pretrained deep learning models. These include visual appearance features from convolutional or vision transformer-based networks, and textual features from models like BERT applied to captions or transcribed speech. Motion-based information such as camera movement and actor pose dynamics are derived using optical flow and pose estimation models, while transitions and scene changes are identified using temporal boundary detection techniques. This multimodal representation enables the model to understand how different forms of stimuli contribute to viewer engagement.

Once the features are obtained, a graph is constructed where each node represents a video segment. Edges in the graph reflect both intra-video and inter-video relationships—such as temporal continuity, shared visual or thematic patterns, hashtags, or engagement metadata like retention rate, likes, and user comments. This structure allows the model to learn from relationships not only within a single video but also across a broader corpus of high-performing content.

Finally, a Graph Attention Network is applied to the constructed graph. This enables the model to focus on the most influential neighboring segments by assigning attention weights based on learned importance. [12] Through this mechanism, the network aggregates relevant relational and multimodal information, producing enriched segment representations. The output of the model is either a hook effectiveness score or a classification label indicating the likelihood of a segment functioning as a strong hook. This framework provides a scalable, interpretable approach for pre-publication content validation based on patterns learned from successful video hooks.

4 Conclusion and Future Work

Understanding the structure and impact of effective video hooks not only serves marketers but also contributes to a deeper understanding of human attention dynamics. Since hooks are designed to trigger immediate cognitive and emotional engagement, analysing their structure provides insight into what captures and retains attention. These findings can extend beyond marketing — for example, informing the design of educational media, where attention is critical for learning. By identifying common patterns that drive engagement, such as visual pacing, emotional tone, or gestural emphasis, this research can support the development of more effective instructional videos, online courses, and digital storytelling tools in education and beyond.

While this study presents a promising methodology for evaluating hook effectiveness using multimodal graph neural networks, future exploration is still needed.

First, the current approach does not include sequence-aware mechanism for better content understanding as content typically has a hook at the beginning, but the narrative tension starts building up as time passes by. Incorporation of temporal graph models or sequence-aware GNNs could capture progression within a video better.

Secondly, the integration of large language models (LLMs) should be explored not only for predicting hook effectiveness, but also for guiding content creators during the creation process. By analysing captions and viewer comments in real time, LLMs could provide actionable feedback and enrich the graph with deeper contextual signals—such as sentiment, common questions, or emotional reactions. This would make the tool more user-friendly and practical, increasing its value and adoption among creators.

Thirdly, another promising direction is to simulate user attention and retention using AI models that mimic how real audiences engage with short-form content. Similar to how AI-powered A/B testing or attention heatmap tools predict viewer focus in UX or advertising, this approach could act as a virtual audience—highlighting which parts of a video are likely to hold attention or cause drop-off. Integrating such simulation into the framework would allow creators to test and refine their hooks pre-publication, reducing the need for trial-and-error on live platforms.

Finally, the ethical dimension of automated attention modeling warrants further investigation. If these models become more effective at predicting and optimising for engagement, future work should consider mechanisms for transparency, fairness, and responsible design, particularly in non-commercial contexts like education or mental health.

5 References

1. Manic, M. (2024). Short-form video content and consumer engagement in digital landscapes. Bulletin of the Transilvania University of Brașov, Series V: Economic Sciences, 17(66), 47–58. https://doi.org/10.31926/but.es.2024.17.66.1.4
2. X. Wang, 2024a; Luo, 2024; Al Haris et al., 2023; Rachmat, Jauhar, et al., 2023
3. Saleem, A., Mehmood, R., Taj, A., Khalid, M. U., Moiz, A., & Lakho, A. (2024). Impact of video content marketing on consumer engagement. Journal of Policy Research, 10(3), 83–95. https://doi.org/10.61506/02.00322
4. Tsapok, O., & Koval, S. (2024). Media's video format of reels on Instagram: Genre and legal aspects of creation. Scientific Notes of the Institute of Journalism, 17(1), Article 12. https://doi.org/10.28925/2524-2644.2024.1712
5. An Observational Narrative of Student Reaction to Video Hooks - MDPI, accessed on April 30, 2025, https://www.mdpi.com/2227-7102/11/6/286
6. Video storytelling ads vs argumentative ads: how hooking viewers enhances consumer engagement - ResearchGate, accessed on April 30, 2025, https://www.researchgate.net/publication/353022418_Video_storytelling_ads_vs_argumentative_ads_how_hooking_viewers_enhances_consumer_engagement
7. Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1301–1309. https://doi.org/10.1109/JSTSP.2017.2764438
8. Sun, C., Baradel, F., Murphy, K., & Schmid, C. (2019). VideoBERT: A joint model for video and language representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 7464–7473). https://openaccess.thecvf.com/content_ICCV_2019/html/Sun_VideoBERT_A_Joint_Model_for_Video_and_Language_Representation_Learning_ICCV_2019_paper.html
9. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks, 20*(1), 61–80. https://doi.org/10.1109/TNN.2008.2005605
10. Wu, L., Sun, P., Fu, Y., Hong, R., Wang, X., & He, X. (2019). SocialGCN: An efficient graph convolutional network based model for social recommendation. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM '19) https://arxiv.org/abs/1811.02815
11. Singh, S. (2024). A comprehensive survey on analyzing social media data for user behavior prediction using graph neural networks. *International Journal of Research Publication and Reviews, 5*(6), 180–184. https://link.springer.com/article/10.1007/s10462-023-10577-2
12. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph attention networks*. arXiv. https://doi.org/10.48550/arXiv.1710.10903

VIDEO:
https://www.awesomescreenshot.com/video/39642953?key=5afa7d0e5cc540903d7371288285a7b1