Artificial Intelligence and Digital Technology

Article

Use Computer Vision and Natural Language Processing to Optimize Advertising and User Behavior Analysis

Qingyang Zhang [1,*]

[1] Simon Business School, University of Rochester, Rochester, NY, 14627, United States
* Correspondence: Qingyang Zhang, Simon Business School, University of Rochester, Rochester, NY, 14627, United States

Abstract: This paper focuses on the technical application of computer vision and natural language processing in automatic advertising placement and user behavior analysis, emphasizing the integration of visual and textual data as the foundation of intelligent decision-making. Building upon the principles of image recognition, semantic understanding, and multimodal content alignment, the study explains how deep learning models extract visual attributes, identify objects and scenes, and interpret text semantics to support precise advertising content generation. By introducing multimodal learning mechanisms, the system can jointly analyze visual cues and linguistic information, thereby enabling richer meaning construction and more accurate situational matching for advertising materials. In this work, an intelligent automatic optimization pipeline is established, covering integrated content generation, personalized customization of advertising materials, dynamic scheduling, and real-time delivery. The framework supports automated creative generation, adaptive content adjustment based on user context, and continuous performance feedback to improve placement strategies. Furthermore, a refined user behavior model is developed by analyzing users' visual interactions-such as gaze patterns, browsing duration, and click distributions-as well as their textual behaviors including search queries, comments, and linguistic preferences. These behavioral signals are embedded into user feature representations to predict potential actions, identify latent interests, and enhance audience segmentation. Overall, the proposed multimodal intelligent advertising system significantly improves ad-matching precision, strengthens behavioral prediction capability, and increases the final conversion rate. The study provides valuable technical references for next-generation intelligent advertising platforms that seek to integrate multimodal perception, adaptive optimization, and user-centric behavioral analytics.

Keywords: computer vision; natural language processing; multimodal learning; advertising placement optimization; user behavior modeling

Published: 13 December 2025

1. Introduction

With the continuous development of the digital advertising ecosystem, traditional advertising release models-characterized by static content design, broad audience targeting, and one-way information dissemination-have become increasingly insufficient to meet consumers' growing demand for personalized, diversified, and context-aware services. As users interact with digital platforms in more dynamic and multimodal ways, advertising systems are required to understand and respond to complex behavioral patterns with higher precision and adaptability [1].

In this context, computer vision (CV) and natural language processing (NLP), representing two core pillars of artificial intelligence, have brought transformative changes to how advertising content is generated, analyzed, and optimized. Computer vision enables machines to interpret visual materials by recognizing objects, scenes, styles,

and emotional cues embedded in images and videos. Meanwhile, NLP supports the interpretation of textual content through semantic parsing, sentiment analysis, topic modeling, and syntactic understanding, allowing deeper comprehension of user intent and content meaning.

More recently, the rapid advancement of multimodal learning has made it possible to jointly process visual and textual information, bridging the gap between images and language [2]. By aligning features across multiple modalities, multimodal models can establish richer semantic associations, thereby enabling more accurate advertising content understanding, creative generation, and relevance matching. This development is particularly valuable in applications such as creative material optimization, content-audience matching, context-aware recommendation, and intelligent placement scheduling [3].

The purpose of this study is to construct an efficient advertising optimization framework that integrates computer vision and natural language processing technologies. Specifically, the research aims to: (1) enhance the accuracy of user interest modeling through multimodal behavioral feature extraction; (2) improve advertising content relevance and delivery efficiency by leveraging visual-textual semantic understanding; and (3) provide technical support and methodological references for the evolution of intelligent digital advertising systems. Through these efforts, the proposed model seeks to address the limitations of traditional approaches and promote the development of more adaptive, intelligent, and user-centered advertising solutions [4].

2. The Core Technical Principles of Computer Vision and Natural Language Processing

2.1. The Basic theory of computer vision in Advertising Content Analysis

The application of computer vision in advertising scenarios mainly focuses on the analysis of element structure, visual style, scene connotation, etc. in advertising images. Common ones include advertising image cascading, visual style analysis, visual positioning and scene recognition, etc. All these are supported by the convolutional neural network (CNN) as the core technology and obtain image features based on the mechanism of local perception. Taking two two-dimensional convolution layers as examples, the output features of the convolutional layer of the neural network can be expressed as:

$$y_{i,j}^{(k)} = f\left(\sum_{m=1}^{M}\sum_{u=1}^{U}\sum_{v=1}^{V} w_{u,v}^{(k,m)} \cdot x_{i+u,j+v}^{(m)} + b^{(k)}\right) \quad (1)$$

Among them, x represents the input feature map, w is the convolution kernel, $b^{(k)}$ is the bias term, and $f(\cdot)$ is the nonlinear activation function (such as ReLU). This method can extract the high-level semantic information of images. Currently, the conventional practice in advertising systems is to use trained models such as ResNet and EfficientNet to extract high-level information such as product elements, trademark logos, and facial features from advertising images.

In addition, object detection models such as YOLO and Faster R-CNN are widely used in the determination of the positions of advertising elements. The system identifies the product type, environment type and scene, etc. through the obvious areas of the image, thereby marking the materials in the advertisement and providing a basis for the next step of material matching and personalized production. Furthermore, there are technologies such as computer vision, which are used in the analysis process of a large number of video advertisements. They extract dynamic features of the video from the changing features within the frame and explore the regions of interest of users on the timeline [5].

2.2. Semantic Modeling Techniques in Natural Language Processing

In the field of advertising, the application of natural language processing technology mainly includes content analysis, sentiment analysis, extraction and generation of advertising texts, as well as the construction of evaluation contents, etc. The purpose of

these tasks is to transform advertising text or customer responses into a series of coherent semantics for subsequent customized information production and recommendation. In recent years, the framework innovation of Transformer has provided technical ideas for NLP models to significantly enhance their capabilities, especially in the aspect of context capture by the Self-Attention mechanism, which has made considerable contributions. Its core calculation is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Among them, Q、K、V respectively represent the mapping matrices of queries, keys, and values, and $\sqrt{d_k}$ is the scaling factor to ensure the stability of the values. This system can guide the model to grasp the emotional tendency, product selling points and user interest preferences when processing advertising copy, and is more suitable for handling advertising statements with different styles and profound semantics.

Because a large number of pre-trained models such as BERT are used, specific adjustments can be attempted to be made to these models for the advertising context (such as the recognition of the purpose of sales promotion titles, product features or emotion-stimulating statements, etc.) As well as the emotional color recognition of evaluation statements and the automatic writing of advertising slogans and A/B testing strategies based on text generation models (such as the GPT series), the speed of advertising innovation generation has been increased several times.

2.3. Multimodal Learning and Cross-Modal Representation Fusion

Due to the intersection of image and text content, advertising media advertisements contain both image and text descriptions simultaneously. In multimodal learning tasks, the semantic representations of text and images are learned in order to achieve semantic matching and joint reasoning to deeply understand the content of advertising materials and improve the reasoning efficiency of the system.

Clustering techniques under the multimodal framework include initial concatenation (concatenating images with text attributes and entering the same framework) and tail concatenation (conducting single-mode training for each module respectively and concatenating at the output stage), etc. The most effective integration technique is the joint embedded learning of the shared semantic space, that is, mapping the image and text representations to the same vector space and then training them in a comparative learning manner.

A typical contrastive learning loss function is InfoNCE:

$$\theta_{\text{InfoNCE}} = -log\frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{N} exp(sim(z_i, z_k)/\tau)} \quad (3)$$

Among them, z_i, z_j represents the feature vector for matching graphic and text pairs, τ is the temperature coefficient, and $sim(\cdot)$ is the similarity measurement function (usually cosine similarity). This method is widely used in pre-trained models such as CLIP and has been successfully applied in scenarios such as image-text retrieval, advertising material matching, and cross-modal recommendation.

After introducing cross-modal learning, the advertising engine can support deep-level interactive operations such as "image-to-text" or "image-to-text", and "image-to-text interaction", further enhancing the intelligence level of the entire system. Especially when modeling user interests, the interaction between text and images also helps to describe the characteristics of user interests more accurately.

3. Intelligent Optimization System Design for Advertising Placement

3.1. Overall System Architecture Design

Accurate, dynamic and individualized advertising placement decisions are the main tasks accomplished by the advertising placement system. In essence, it is a multi-step data analysis and optimization decision-making sequence. The system design needs to cover

all the links from the beginning to the end of advertising placement, from the initial collection of materials, content analysis, establishment of user interaction models to strategic decision-making, forming a circular feedback. The system collects different types of information from the advertisers' side and the platform's perspective, including pictures, text or video materials, etc., as well as user action trajectories such as clicks, readings, dwell times, and consumption. In addition, in the content parsing stage, pre-trained visual and language models are applied to the original material advertisements to generate contents such as product categories, style tags, keywords, and emotional attributes, which are stored in a structured and formal manner. In the user modeling stage, a multimodal fusion network is trained based on previous behavior data to form a deeply complex user interest vector. Form users' viewpoints, semantic preferences, and behavioral trajectories, and use them as important bases for advertising matching and strategy optimization.

In the system architecture, the strategy part integrates user portraits, characteristic representations of advertising materials, and context information (such as placement time, terminal type, positioning latitude and longitude, etc.), uses CTR/CVR prediction models or reinforcement learning models to output the optimal set of advertising candidates, and combines placement restriction rules and resource scheduling to make specific placement decisions. Finally, the placement data is processed through a feedback mechanism to obtain feedback indicators (such as click-through rate, conversion rate, exposure duration, etc.), which are used to optimize the model and material strategy, achieving a data-driven closed-loop optimization mechanism. The following Figure 1 is the entire process logic diagram of the system

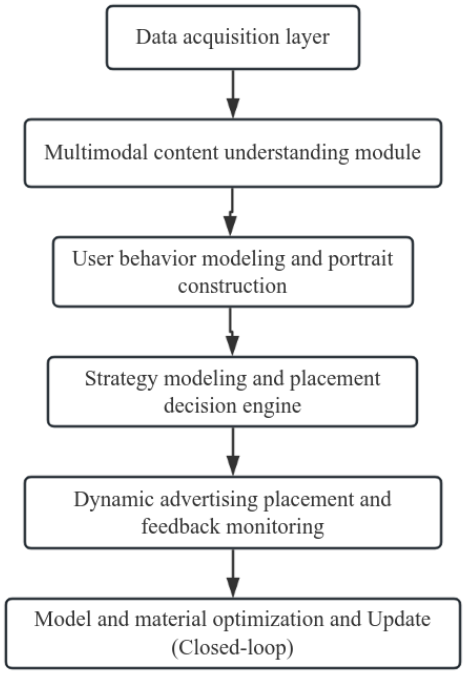

Figure 1. Overall system architecture.

Each module is expressed in a modular modeling manner and connected by decentralized information flows, providing the possibility for efficient parallel processing and real-time response of the entire system. At the same time, it is also conducive to the hierarchical optimization design of offline training and online reasoning.

3.2. Optimization and Personalized Generation of Advertising Materials

The core of optimizing advertising materials lies in the degree of alignment with the needs of the target customers. For images, generative adversarial networks such as StyleGAN are used to adjust the patterns to achieve better appeal, or to remove the existing noise and make structural adjustments. For text content, tools such as GPT and T5 are invoked to convert advertising text and ensure the smoothness and fluency of the text semantics. To improve the generated results, the Top-k extraction method or Beam Search decoding is adopted, and the vocabulary reordering strategy is applied to ensure semantic consistency. The quality of the generated materials is evaluated using a multi-mode scoring network. The pattern scoring network is usually a small neural network, which is used to receive various visual elements and output the estimated scores of click-through rate and conversion rate to screen out the advertisement content with good estimated effects, establish a closed loop of "generation - screening - feedback", and ultimately achieve the goal of comprehensive efficiency improvement.

3.3. Modeling and Scheduling Mechanism of Placement Strategy

Based on the prediction of CTR/CVR, the placement strategy model is further constructed, taking user behavior sequences, advertising content encoding and advertising background information as inputs. For example, the DIN model uses the method of local attention to simulate the changes of user interests over time to improve the accuracy of matching. Output the serialized advertisement candidate set and carry out the placement under budget constraints. Furthermore, reinforcement learning is used to model the placement as an MDP problem. There are mainly two ways to use strategy learning. One is to learn strategies using strategy gradient methods such as PPO, with the aim of maximizing long-term conversions or LTV and other benefits. The application of this strategy usually requires budget constraints and a long advertising interaction time. Another type of strategy learning method is causal reasoning, which uses IPS and double robust estimators to correct the bias of the strategy estimator. Meanwhile, multi-arm gambling algorithms, such as Thompson Sampling, are also used to continue exploring the effects of other materials without overly affecting the main effect. After the strategy is determined, the scheduler is used to implement the effects such as frequency control, budget allocation, and exposure ceiling.

4. User Behavior Analysis and Interest Modeling

4.1. User Visual Behavior Analysis

For users of advertising platforms, visual activities such as browsing, clicking, swiping, gazing, and legend interaction are all characteristics that show interest and inclination towards a certain advertising image or text. By tracking these behaviors and combining them with the image features of advertising elements, the browsing and selection features of users for various advertising pictures can be formed, thereby improving the personalized recommendation performance. These visual activity data usually come from the front-end insertion points of the client side, the server-side logs, and the visual area interaction processing devices (such as mouse path heat maps, video picture eye predictions, etc.). After preprocessing, it becomes analyzable data input. The following Table 1 lists the common data fields in visual behavior modeling and their corresponding meanings:

Table 1. Key Fields and Descriptions in User Visual Behavior Modeling.

Field name	Data type	Explanation
User ID	String	User unique identifier, used for behavior attribution
Material ID	String	The identification of the advertisement image/video material being watched

Duration of stay(ms)	Integer	The duration that users spend on advertising materials reflects the intensity of their interest
Click behavior(0/1)	Boolean type	Whether the user clicks on the advertisement material
Sliding rate(Pixels /s)	Floating-point number	The speed of page swiping indirectly reflects whether the content is quickly skipped over
Scaling behavior(0/1)	Boolean type	Whether to perform zooming and zooming operations on the image
Visual style label	Multi-classification label	Image style recognition results, such as "minimalist", "passionate", "tech-savvy", etc
Object recognition result	Multiple labels	The core object categories that appear in the image, such as "people", "clothing", and "products"
Timestamp	Time type	The time point at which the behavior occurs is used for sequence modeling and time segmentation analysis

After being actually put into use, this system will provide feedback based on the user's "preferences" for these resources in each field, establishing visual preference characteristics for this user. For instance, if a certain user always opts for "minimalist design" advertisements and maintains a long viewing record of them, then he may have a high interest in categories with distinct technological styles and minimalist designs.

In addition, the platform can also adopt the content analysis method to extract the advanced meaning attributes of the pictures (such as color pattern, background structure and emotional hierarchy, etc.), and use this as a standard to enrich the multi-level information volume, so as to achieve the purpose of more comprehensively constructing the user's visual preferences. By recording the formatted behavior and marking the content of the attached material advertisements, the platform can quickly outline the user's visual preferences in a very short time, and achieve differentiated matching and dynamic correction in the subsequent material push.

4.2. Modeling of User Text Behavior

In addition to visual features, the text of user behavior may also have implicit demands, such as query terms, the content of evaluations, inquiries about products, and the text shared about products on social networks, etc. Through the methods of natural language processing, valuable words or phrases are mined from them, and models are established to understand language habits and thought intentions. For the already cleaned data, query words, labels, word vectors or BERT encoders can be used to generate low-dimensional vector representations. Conduct sentiment, goal and entity extraction analysis on unstructured text (such as comments or question-and-answer text) to discover which aspect of the brand/product function/service the user focuses on.

Let a set of text behavior sequences generated by the user be $\{T_1, T_2, \ldots, T_n\}$, and the semantic representation of each text be t_i. Then the semantic interest representation of the user can be modeled in the form of attention-weighted aggregation:

$$T_{\text{user}} = \sum_{i=1}^{n} \alpha_i \cdot t_i, \alpha_i = \frac{exp(q^T t_i)}{\sum_{j=1}^{n} exp(q^T t_j)} \tag{4}$$

Among them, q is the context query vector, and α_i is the attention weight of the i-th behavior. This system can automatically capture the user's most recent gaze target, thereby improving the adaptability and accuracy of semantic modeling.

4.3. User Profile Construction and Behavior Prediction

After establishing the user's visual and text interaction mode, the advertising platform will combine several forms of interest features into a unified user portrait to guide the subsequent click probability estimation, conversion rate estimation and

personalized decision-making. A user's portrait consists of both basic static attributes such as gender, age, region, and hardware configuration, as well as higher-level dynamic behavioral attributes. Visual preference values and the content features of textual information, etc. are all important components of multi-feature representation, and they should all be uniformly processed in the meaning space. Therefore, deep learning techniques are usually adopted for them to reconstruct and fuse, generating high-level user representation vectors. This vector can be directly input into the CTR/CVR prediction model or the recommendation ranking algorithm to achieve personalized decision-making.

In behavior prediction, the fused user characteristics and a series of advertising feature representations that can be used for placement are sent into the model. Based on the model, relevant values such as whether it can arouse a certain user's attention to the advertisement, browsing duration and conversion probability are predicted. Sequence modeling such as Transformer and GRU, as well as attention methods for modeling, are adopted. This attention is taken as the user's intention to predict the changes in user attention over time and behavior, as well as the long-term dependence between behaviors. This predicted value can also be applied to the reordering of recommendation ranking and the strategy adjustment of advertisement release, which is a closed-loop process of iteration from behavior to interest, then to placement to behavior feedback (see Figure 2).

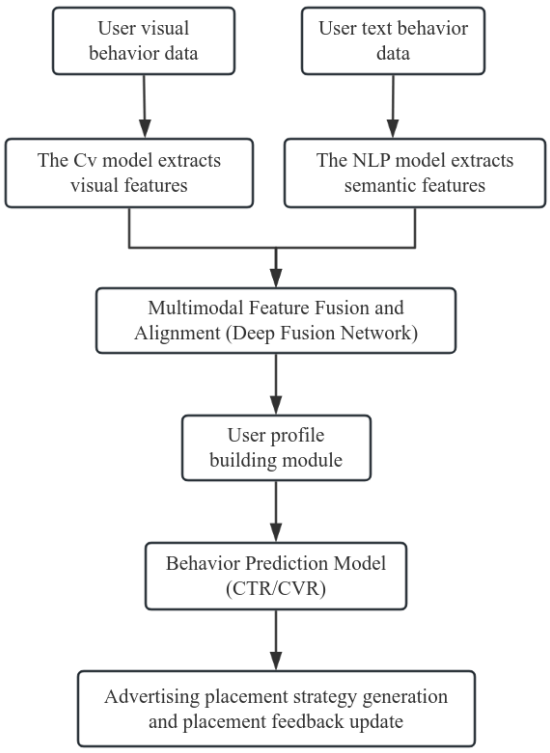

Figure 2. Flowchart of User Profiling and Behavior Prediction.

Through the above process, the multimodal user behavior is recognized and predicted, and the user profile and advertising strategy model are evaluated. The feedback from users is continuously corrected. On the one hand, it enhances the personalization of advertising accuracy; on the other hand, it effectively increases the conversion rate of the entire advertisement.

5. Conclusion

This paper investigates advertising placement strategies and the construction of user patterns by systematically integrating computer vision and natural language processing technologies. Through a comprehensive analysis of visual and textual content processing methods, the study proposes an intelligent advertising architecture capable of multimodal content understanding, cross-modal feature alignment, and personalized strategy generation. The framework not only enhances the semantic interpretation of advertising materials but also supports dynamic content adaptation based on user context and behavioral cues.

By leveraging visual semantic construction, multimodal representation learning, and predictive modeling of user behaviors, the proposed approach significantly improves the accuracy of advertising adaptation. It enhances audience engagement by matching content more closely with user preferences and behavioral characteristics, ultimately contributing to higher relevance and conversion performance. Moreover, the study outlines a feasible and scalable intelligent deployment route, providing practical guidance for implementing multimodal-driven advertising systems in real-world scenarios.

Overall, the findings demonstrate that integrating CV and NLP into advertising decision-making processes can effectively strengthen content understanding, refine user modeling, and optimize placement strategies. This work offers both theoretical and technical references for the development of next-generation intelligent advertising platforms that emphasize personalization, adaptability, and data-driven optimization.

References

1. S. M. Andersen, S. Chen, and R. Miranda, "Significant others and the self," *Self and identity*, vol. 1, no. 2, pp. 159-168, 2002. doi: 10.1080/152988602317319348
2. R. Kapuscinski, "The other," *Verso Books*, 2018.
3. M. K. F. Yip, and V. W. Y. Lum, "Beyond images: data visualization through headline analysis in historical newspaper with computer vision," In *International Workshop on Signal Processing and Machine Learning (WSPML 2023)*, December, 2023, pp. 279-285.
4. S. Talafha, "Generative Models in Natural Language Processing and Computer Vision," *Southern Illinois University at Carbondale*, 2022.
5. J. Dong, "Natural Language Processing Pretraining Language Model for Computer Intelligent Recognition Technology," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 8, pp. 1-12, 2024. doi: 10.1145/3605210