International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 7, Number 2, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v7n2.03
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research on Cross-Modal Interaction Techniques between Natural Language Processing and Computer Vision

Shuo Song

Fakulti Teknologi dan Sains Maklumat, Universitiy Kebangsaan Malaysia, Malaysia

ABSTRACT

With the penetration of artificial intelligence technologies into multi-scenario applications, single-modality technologies are no longer able to meet the demands of complex tasks. While NLP can parse text semantics, it lacks the intuitiveness of visual information; while CV can process image pixel features, it struggles to understand the abstract instructions conveyed by text. Against this backdrop, cross-modal interaction techniques between NLP and CV have become a key approach to overcoming these bottlenecks. This paper examines the core logic of cross-modal interaction, first clarifying the essential characteristics and interaction goals of modal heterogeneity. It then analyzes key techniques for extracting modal representations, and then explores implementation paths for cross-modal alignment (semantic matching and spatial mapping) and fusion (at the feature, semantic, and decision levels). The effectiveness of these techniques is validated using real-world application scenarios such as visual question answering (VQA) and image captioning. Finally, the paper summarizes current challenges, such as modality imbalance and insufficient robustness, and proposes optimization strategies that combine knowledge graphs with lightweight models. Research indicates that efficient cross-modal interaction requires "precise alignment" as its foundation and "deep fusion" as its core. The implementation of these technologies can significantly enhance the perception and decision-making capabilities of AI systems in complex environments, providing technical support for fields such as intelligent human-computer interaction and autonomous driving.

KEYWORDS

Natural Language Processing; Computer Vision; Cross-Modal Interaction; Modal Alignment; Feature Fusion; Visual Question Answering

1. INTRODUCTION

The early development of artificial intelligence focused on single-modal technologies: In the natural language processing (NLP) field, pre-trained models enabled deep semantic parsing of text, enabling tasks such as machine translation and sentiment analysis; in the computer vision (CV) field, convolutional neural networks (CNNs) and visual transformers (ViTs) enabled image classification and object detection. However, in real-world applications, human cognition often relies on multimodal information, combining "text + vision." For example, drivers need to simultaneously understand navigation instructions and visual road scenes, while intelligent customer service agents must combine user text descriptions with uploaded image fault information. The limitations of single-modal technologies have become increasingly apparent: NLP alone cannot determine the state of objects in images, and CV alone struggles to understand abstract requirements in text. The core value of cross-modal interaction technology lies in breaking down the modal barriers between natural language processing (NLP) and computer vision (CV), enabling the complementarity and synergy of semantic and visual information. In recent years, this technology has demonstrated its potential in a variety of fields: image captioning can automatically generate semantically coherent captions for

images, and visual question answering can answer text-based questions based on image content. These applications rely on the deep interaction between natural language processing (NLP) and CV. However, current research still faces numerous challenges: modal heterogeneity (text is a discrete sequence, images are continuous pixels, and the data structure differs significantly), making direct feature matching difficult; interaction robustness in complex scenarios; and limited model generalization in scenarios with small sample data. Based on this, this article systematically analyzes the key technologies of cross-modal interaction between NLP and CV from three perspectives: basic theory, technical approaches, and application scenarios. It also identifies current challenges and proposes optimization directions. The aim is to provide a reference for related research and engineering applications, and to promote cross-modal technology from the laboratory to practical scenarios.

2. BASIC THEORY AND CORE OBJECTIVES OF CROSS-MODAL INTERACTION

The essence of cross-modal interaction technology is to address the "information gap" between natural language processing (NLP) and computer vision (CV). Text modalities convey abstract semantics through character sequences, while visual modalities present concrete features through pixel matrices. These two modalities exhibit significant differences in their data sources, structures, and expression logic. This "modal heterogeneity" is the primary obstacle that interaction technology must overcome. From a theoretical perspective, cross-modal interaction requires establishing a "semantic-visual" mapping relationship, enabling the comparability, complementarity, and synergy of information from both modalities within a unified space. The core objectives of cross-modal interaction can be categorized into three key areas: First, semantic consistency, meaning that the meaning of the textual expression matches the visual content presented. For example, the text "a puppy running on the grass" must correspond to the objects "puppy" and "grass" in the image, as well as the action "running." Second, information complementarity, where the information from both modalities is combined to compensate for the shortcomings of a single modality. For example, the text "broken window" can help the CV locate the damaged area of the "window" in the image, while the detailed features of the image can supplement the "extent of damage" that the text does not mention [1]. Third, task adaptability, where interaction technology must adapt the modal collaboration method to the specific task. For example, in visual question answering, the CV should first locate the target, followed by the NLP interpreting the question and generating an answer based on visual features. From a technical perspective, cross-modal interaction can be broken down into three key steps: modal representation, modal alignment, and modal fusion. Modal representation is the foundation, requiring the conversion of text and image into machine-understandable vector features; modal alignment is the core, requiring the establishment of a precise mapping between the two modalities; and modal fusion is the goal, integrating the aligned information into a unified decision-making basis. These three steps are interrelated. If the feature quality during the representation phase is insufficient, alignment accuracy will be directly affected; if alignment is biased, the fusion result will deviate from the task requirements.

3. MODAL REPRESENTATION AND FEATURE EXTRACTION TECHNOLOGY

Modal representation is the first step in cross-modal interaction. It requires converting text and images into structured feature vectors, each of which must preserve the core information of the modality (the semantics of the text and the visual attributes of the image). The feature extraction logic of different modalities varies significantly, and an adaptive approach must be selected based on the respective technical systems. Pre-trained models have become the mainstream technology for text modality

representation. Models such as BERT use a "bidirectional attention mechanism" to capture the contextual semantics of text. For example, when processing "a cat sits on the sofa," they can identify the positional relationship between "cat" and "sofa." GPT models, on the other hand, optimize the semantic coherence of text through "autoregressive generation," making them more suitable for generation tasks. The advantage of these models is that they acquire universal semantic representation capabilities through large-scale pre-training on unlabeled text, allowing them to be adapted to cross-modal scenarios with only a small amount of fine-tuning on task data [2]. Furthermore, text representation must strike a balance between fine-grainedness and simplicity. Excessively fine-grained features increase computational complexity, while overly coarse features lose crucial information. Currently, a "word embedding + sentence-level aggregation" approach is commonly used to balance these two aspects. For example, word vectors are combined into sentence vectors through average pooling. Regarding image modality representation, the technical path has evolved from focusing on local features to focusing on global semantics. Early CNNs extracted local features from images through convolutional layers, but struggled to capture global semantic connections. The ViT model, which incorporates the Transformer attention mechanism, segments images into "tiles" and converts them into sequences. This allows it to simultaneously focus on local details and global layout. For example, when recognizing an "apple on a table," it can capture both the color of the apple and contextual information about the table. Furthermore, image representation must address scale differences. The size and angle of the same object may vary across images, requiring data augmentation and adaptive pooling techniques to ensure feature vector consistency. To achieve a fundamental condition for cross-modal interaction, the representations of both modalities must be spatially compatible—that is, text features and image features must be mapped into a vector space of the same dimensionality. Currently, contrastive learning is commonly used to achieve this goal. For example, the CLIP model, through large-scale training on image-text pairs, brings semantically similar text and images closer together in vector space, laying the foundation for subsequent alignment and fusion.

4. IMPLEMENTATION PATH OF CROSS-MODAL ALIGNMENT TECHNOLOGY

Cross-modal alignment is a core technology for addressing modal heterogeneity. Its goal is to establish a precise mapping between text and image—including both semantic and spatial alignment. Different alignment requirements correspond to different technical paths, and the appropriate solution should be selected based on the task scenario. The core of semantic alignment is to match text semantics with image content. Commonly used techniques include attention mechanisms and semantic mapping. In terms of attention mechanism, "cross attention" is used to make text and image pay attention to each other's key information. For example, in the visual question-answering task, when the question is "How many red birds are there in the picture", the model will let the "red" and "bird" keywords in the text focus on the area of red birds in the image and ignore irrelevant background; in terms of semantic mapping, by establishing a "text concept-image category" correspondence, for example, mapping "transportation" in the text to "car", "airplane" and other target categories in the image, this mapping can be trained through labeled data, and can also be supplemented with prior knowledge with the help of knowledge graphs. In addition, for scenarios with unlabeled data, self-supervised semantic alignment technology is gradually emerging. For example, through the "image clustering + text clustering" method, similar images are automatically matched with similar texts, reducing dependence on labeled data. Spatial alignment is mainly suitable for tasks that require locating specific areas of the image. The core is to achieve the correspondence between "text description location" and "image spatial area". Common methods include "region-level alignment" and "pixel-level alignment." Region-level alignment, based on object detection technology, first segments the image into multiple target regions and then matches noun phrases in the text to these regions. For example, the text "the tree on the left" corresponds to the tree region on

the left side of the image. Pixel-level alignment is more refined, using a segmentation model to obtain a pixel mask of the target and then matching the attributes of the text description to the corresponding pixel features. Spatial alignment is challenging in handling "ambiguous descriptions." For example, the location of the text "the distant mountain" is unclear. In this case, alignment requires combining image depth information with scene common sense. Current alignment technology is limited in its adaptability to complex scenarios [3]. For example, alignment accuracy significantly decreases when there are occlusions in the image or ambiguous text. Future efforts will require integrating contextual reasoning capabilities to improve alignment robustness.

5. CROSS-MODAL FUSION STRATEGIES AND APPLICATION SCENARIOS

Cross-modal fusion integrates aligned text and image information to form a unified decision-making basis. The strategy should be selected based on task requirements, with different fusion stages corresponding to different technical logic. Furthermore, the implementation of fusion technology must be based on specific application scenarios, with actual needs informing technical optimization. Based on the fusion stage, cross-modal fusion can be divided into "early fusion," "mid-term fusion," and "late fusion." Early fusion directly integrates the raw features of text and image, for example, concatenating text and image vectors before inputting them into a neural network. This approach preserves fine-grained information and is suitable for tasks requiring detailed support, such as image captioning. However, it places high demands on feature quality, and if the features of the two modalities differ significantly, this can easily lead to "fusion redundancy." Mid-term fusion integrates the semantics after modal alignment. For example, through the cross-attention of the Transformer, text and image semantics are mutually reinforced, preserving key information while reducing redundancy [4]. This approach is currently the mainstream approach for visual question answering (VQA). Late fusion first allows NLP and CV to generate separate decision results, then integrates them through weighted voting, logical reasoning, and other methods. This approach offers robustness and is suitable for safety-critical scenarios such as autonomous driving. For example, a turn is only executed when a textual instruction "turn left" agrees with a visual inspection decision of "no obstacle on the left." Cross-modal fusion technology has been implemented in multiple fields. First, image captioning: CV extracts object and scene features from an image, and NLP combines these features to generate a semantically coherent text description. For example, generating a sentence for an image of "children flying kites in a park" requires the synergy of early fusion and semantic alignment. Second, visual question answering (VQA): Users pose text questions, and the model generates answers based on image content. For example, for the question "What is the color of the bicycle in the picture?", mid-term fusion is required to achieve a deep interaction between the "question semantics" and "image color features." Third, autonomous driving environment understanding: Late-term fusion of text instructions and visual images enables integrated decision-making to ensure driving safety. Fourth, intelligent customer service fault diagnosis: Feature-level fusion of user text descriptions and uploaded fault images allows for rapid identification of fault types and solutions. The practical application of these scenarios demonstrates that the choice of fusion strategy must be aligned with the task objectives: early fusion is preferred for tasks requiring detailed information, mid-term fusion for tasks requiring semantic coordination, and late fusion for tasks requiring high robustness.

6. CURRENT CHALLENGES AND OPTIMIZATION DIRECTIONS

Although cross-modal interaction technology for NLP and CV has made significant progress, practical applications still face numerous challenges. These challenges stem from both the limitations of the technology itself and the varying requirements of complex scenarios. Identifying these challenges and proposing optimization directions are key to promoting the practical application of

this technology. Current technologies face three core challenges. First, modality imbalance: The information density of images is much higher than that of text, leading to "visual information overload" or "text information deficiency" during fusion. For example, in visual question answering, models may over-rely on image details and ignore the core requirements of text questions. Second, insufficient robustness in complex scenarios: When images are occluded or blurred, or when text contains ambiguities or spelling errors, the accuracy of modal representation and alignment decreases significantly, compromising fusion results. Third, weak small-sample learning capabilities: Existing cross-modal technologies often rely on large-scale annotated image-text pairs. However, in small-sample scenarios such as medical diagnosis and industrial testing, models generally generalize poorly, making them difficult to implement. To address these challenges, three optimization directions can be explored. First, balance the weight of modal information: Through "attention weight adjustment" and "information filtering mechanisms," we reduce the interference of redundant visual information and enhance the influence of key textual information. For example, during fusion, we assign higher weight to core keywords in the text and lower weight to irrelevant background areas in the image. Second, improve robustness in complex scenarios: Introduce "modal completion technology," such as using image inpainting models to restore details in blurred images and text error correction models to correct spelling errors. Combined with "multi-scenario adaptive training," we train models on datasets containing occlusion and noise to enhance their robustness. Third, overcome the bottleneck of small-sample learning: Combining "transfer learning" with "knowledge graphs," we transfer model parameters trained on large-scale general data to small-sample scenarios. Knowledge graphs are then used to supplement domain knowledge, reducing reliance on labeled data. Self-supervised learning techniques can also reduce training costs for small-sample scenarios. Furthermore, lightweight optimization is a key area [5]. Current cross-modal models often rely on complex networks, making them difficult to deploy in resource-constrained scenarios such as mobile phones and edge devices. Future efforts will require model compression and efficient network architecture design to reduce computational and storage overhead while ensuring performance.

7. CONCLUSION

This paper examines cross-modal interaction technologies between natural language processing (NLP) and computer vision (CV), systematically analyzing the core components and practical approaches of the technology ecosystem. This research demonstrates that the essence of cross-modal interaction technology lies in breaking down barriers to modal heterogeneity and achieving complementarity between textual semantics and visual information through the collaborative process of "modal representation, alignment, and fusion." At the modal representation level, models such as BERT and ViT provide effective solutions for extracting structured features from text and images, while contrastive learning achieves spatial compatibility between the two modalities. At the modal alignment level, semantic alignment and spatial alignment address the "content matching" and "position correspondence" problems, respectively. At the modal fusion level, early, mid, and late fusion strategies must be adapted to task requirements, with practical applications already covering image description, visual image processing, and image processing. This research also highlights the limitations of current technology: modal imbalance leading to information weight imbalance, insufficient robustness in complex scenarios, and weak small-sample learning capabilities. These issues remain key constraints to its practical application. Future optimization efforts should focus on three key areas: balancing modal weights, improving anti-interference capabilities, and overcoming the small-sample bottleneck. Simultaneously, lightweight design should be incorporated to facilitate model deployment on edge devices. The value of this paper lies in providing a comprehensive perspective on cross-modal interaction technology from theory to application. The technical paths and challenges it outlines can serve as a reference for research and engineering practice in related fields. It should be noted that this paper does not conduct in-depth experimental validation of a single fusion strategy. Subsequent quantitative research targeting specific application scenarios could

further enhance the relevance of the technical analysis. With the advancement of large-scale model technology and the accumulation of multimodal data, cross-modal interaction between NLP and CV will become more efficient and robust, potentially playing a greater role in intelligent human-computer interaction, smart healthcare, autonomous driving, and other fields, driving AI from "single-modal perception" to "multimodal cognition."

REFERENCES

[1] Li Xu, Zhu Rui, Chen Xiaolei, et al. A review of hallucinations in large visual language models: causes, evaluation and governance [J/OL]. Computer Research and Development, 1-24 [2025-09-02]. https://link.cnki.net/urlid/11.1777.TP.20250506.1509.006

[2] Jiang Xiurong. Research on salient object detection algorithm based on multimodal information fusion [D]. Beijing University of Posts and Telecommunications, 2024. DOI: 10.26969/d.cnki.gbydu.2024.000132.

[3] Huang Yupan. 1. Research on multimodal intelligence for vision and language representation learning [D]. Sun Yat-sen University, 2023. DOI:10.27664/d.cnki.gzsdu.2023.000017.

[4] Wu Siying. Research on cross-modal semantic alignment method for vision and language [D]. University of Science and Technology of China, 2023. DOI:10.27517/d.cnki.gzkju.2023.000627.

[5] Zhang Ran, Wang Lei, Gao Xiangyi, et al. Research on the application of multimodal intelligent interaction technology in digital banking [J]. China Financial Computer, 2024, (02): 34-36.