Image and Text Fusion for Intent Detection in Multimedia

Wang Meng

Abstract

In the field of human-computer interaction, human intention detection is a challenging problem and a key link to achieving barrier-free communication between humans and machines. With the rapid evolution of multimedia technology and the widespread use of social media platforms, the detection of user intention has become increasingly challenging. Traditional unimodal approaches, especially those relying solely on either textual or visual information, often fall short of capturing the intricacies of user intentions in multimedia content. To address this limitation, the fusion of image and text modalities using multimodal technology has emerged as a promising solution for intention detection.

Compared with single-modal data such as images and text, multimodal data can contain more information and can more accurately identify user intentions. At present, there are few studies on intention detection based on image and text fusion, and they mainly consider how to integrate modal features in the feature fusion stage, while in the feature extraction stage, only a single-modality pre-trained model is used, which not only increases the complexity of the system, which increases the computational cost and may also limit the model's comprehensive understanding of the global context, making it difficult to effectively capture the correlation information between modalities.

This research endeavors to design a new intention detection framework, which includes two equally important stages of multimodal representation and fusion, to explore the integration of image and text data to enhance the accuracy and robustness of intention detection in multimedia content. In the feature representation part, the CLIP multimodal large-scale pre-training model is used to simultaneously extract text and image features, which simplifies system integration and saves computing resources while learning the associated information between modalities. In the feature fusion part, due to the different importance of text and pictures, an attention-based cross-modal fusion method is designed, which enables the model to dynamically adjust the attention to different modalities at each moment and capture important features, while reducing

noise. Finally, to verify the effectiveness of the proposed model, this study developed an intention detection framework and carried out experimental verification on the intention detection dataset.

Key Words: intention detection, multimodal technology, CLIP, feature representation, multimodal fusion

Table of Contents

Chapter-1: Introduction

1.1. Background

As an important research direction in the fields of artificial intelligence and computer science, intention detection aims to enable computer systems to understand and interpret users' true intentions, thereby responding to user needs more intelligently. In recent years, with the rapid development of multimedia technology, the forms of information released by users have become more diverse. When many users publish text information, they usually add corresponding picture information to express their true intentions more vividly and intuitively. This kind of Information in the form of multimedia better meets the needs of users to express themselves, obtain information, and participate in interactions on social media. It also brings new challenges and opportunities to intention detection.

In recent years, machine learning has made remarkable progress in processing various forms of media such as images and texts. Especially, the wide application of deep learning technology provides a powerful tool for intention detection, so that the model can better learn and understand the real purpose of users from complex massive data. The application of these technologies promotes the progress of intention detection technology in various fields and provides a more accurate and intelligent user interaction experience for intelligent systems. However, traditional intention detection methods are usually limited to single-modal data analysis and do not make full use of the rich information of multimedia data. In practical applications, users often use multiple forms of communication such as images and texts at the same time, and single-modal processing methods are difficult to fully understand and recognize the user's real intention.

In the era of digital multimedia, the field of intention detection is facing more complex and diverse user expressions. As a cutting-edge research method, multimodal data fusion, especially the fusion of image and text, provides a new idea to solve this

problem, as shown in Fig.1.1. It is generally believed that multimodality usually mainly contains three parts: representation, alignment, and fusion [1], among which multimodal representation and fusion are more important and are usually the focus of research [2]. Inspired by the single-modal pre-training model, a series of multimodal pre-training works based on image-text data have emerged. Multimodal representation pre-training aims to use self-supervised learning paradigms, including contrastive learning, mask self-supervision, etc., to train on large-scale image-text correlation data. The model obtains universal and strong generalization ability of multimodal representation. Compared with a single modality, multimodal fusion contains more information. On the one hand, this information can complement each other. On the other hand, there is also the possibility of false data or contradictory phenomena. Therefore, how to correctly use and deal with the relationship and contradiction between each modality to improve the accuracy and robustness of intention detection has become one of the main research directions in this field.

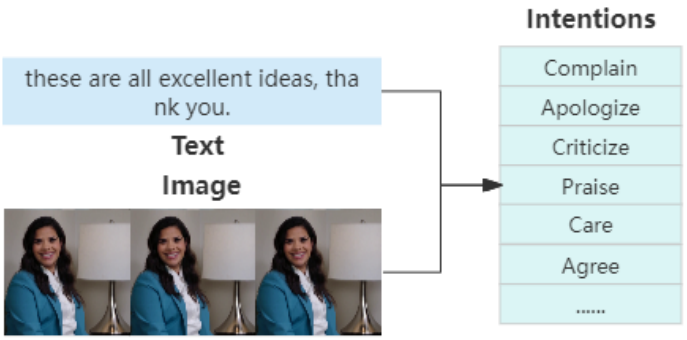

Fig.1.1. An example of intent detection based on text and image fusion.

1.2. Motivation

Multimedia includes text, image, audio, video, and other forms of information expression. In recent years, the explosive growth of social media platforms has provided a wide range of dissemination platforms for multimedia content, while multimedia forms of information better meet the needs of users to express themselves,

obtain information, and engage in interaction on social media, but also bring challenges, especially in understanding and predicting user intentions in these digital Spaces. As one of the core technologies of human-computer interaction, intention detection aims to accurately identify the user's intention from his input to achieve more intelligent services. In multimedia technology, understanding whether a user's comment is truly informative, ironic, supportive, or critical through intention detection is crucial for improving recommendation systems, content moderation, and fostering healthier online communities. However, there are some potential weaknesses in traditional intention detection methods, such as only single modal data such as text or image are often considered, and the advantages of multimedia data are not fully utilized. Existing research based on image-text information fusion mainly focuses on sentiment analysis and content classification, but there are still significant gaps in interpreting the subtle intentions behind user behavior. Based on this, this study proposes an intention detection method based on image and text fusion. Through this research, we can not only verify the feasibility and effectiveness of image and text information fusion technology in intention detection, improve the accuracy of intention detection in multimedia environments, but also make it more suitable for diverse user communication scenarios in the real world.

1.3. Problem Statement

The traditional intention detection method is mainly based on manual rules and infers the user's intention through a pre-defined rule set. Although this method is simple, it requires manual arrangement of the rule set, has limited coverage and is not flexible enough. With the development of deep learning technology, researchers have proposed many intention detection methods based on deep learning [3, 4]. These methods can automatically learn relevant patterns from massive data and are more efficient and accurate than manual rules. However, using only a single modality (for example, text modality) for intention detection often cannot fully utilize the diversity of information, and it is difficult to effectively complete the task in the face of the diversity and

complexity of input content. Therefore, how to effectively integrate and utilize multimodal data in a multimedia environment is one of the main research issues in this field.

Nowadays, various multimodal technology related solutions and methodologies have emerged to address the above problem in intention detection, for instance:

In [5], a model is proposed to capture the complex meaning multiplication relationship between image and text in multimodal Instagram posts. While this model integrated text and image information to identify the intention, in the multimodal fusion stage, only the simple fusion strategy is used, that is, adds the two vectors, and the interaction information between text and images is ignored.

In [6] introduces a late-fusion approach for integration of the video signal with the captions signal for intention Detection. Although it shows significant improvements with unimodal pre-trained models, the HERO used in the article is a pre-trained model only for video language, heterogeneity between modalities is not considered and its performance on image and text data is not yet known.

In [7] develops a novel feature fusion method based on attention mechanism, which can recognize and utilize the importance of different modalities. It designs complex strategies in feature fusion to reduce possible noise but uses original pre-trained models Bert and ResNet50 to extract text and image features respectively in feature extraction, which limits the model's comprehensive understanding of the global context.

In response to these challenges of insufficient single-modal detection capability and the limitations evident in the aforementioned multimodal methods, this research endeavors to design a new intention detection framework, which includes two equally important stages of multimodal representation and fusion. In the feature representation stage, we use a multimodal large-scale pre-trained model to extract image and text features and achieve multimodal representations, and in the fusion stage, since the importance of text and pictures is not the same, we design a cross modality fusion method based on attention, this research aims to bolster the overall effect of intention

detection.

1.4. Research Significance

In view of the large amount of image and text data existing in multimedia platforms, this study has important research significance for accurately understanding and identifying the true intentions of users by integrating image and text information. First, the intention detection of image and text integration has important application value in social media analysis. By analyzing the intention of posts containing image and text information, we can understand the real thoughts of users, thereby providing personalized services such as advertising marketing and public opinion monitoring. Secondly, many intelligent products are already well known. By integrating information from different media for more accurate intent detection, intelligent systems can be provided with a more flexible and natural interaction method, thereby making correct judgments or reasonable feedback on human intentions to improve the quality and efficiency of human-computer interaction. Finally, in actual application scenarios, users often communicate in diverse ways, and intent detection based on image and text data can better adapt to this diversity, thereby improving the robustness and applicability of the intention detection system. Intention detection based on image-text fusion has become an important research direction in the field of artificial intelligence. This trend not only promotes the development of human-computer interaction technology, but also provides more powerful and intelligent application prospects for intelligent systems in various fields.

1.5. Research Questions

1. How can the multimodal large-scale Pretrained models be leveraged for feature extraction and representation in the field of intention detection?
2. How to develop the proposed intention detection framework?
3. How can the proposed framework affect the accuracy of intention detection?

1.6. Research Objectives

1. To introduce multimodal large-scale pre-training models to extract text and image features to achieve multimodal representation.
2. To develop the proposed intention detection framework based on image and text fusion.
3. To evaluate the performance of the proposed intention detection framework by comparing its accuracy with the baseline model.

Chapter-2: Literature Review

2.1. Intention Detection

Recently, with the rapid development and application of multimedia technology, users are now more inclined to express their intentions through multimodal data such as text and images. In fact, multimodal data contains richer information, and the accuracy of intention detection can be improved by learning from multimodal data. At present, intention detection based on image and text fusion has become a research hotspot in the field of artificial intelligence. In the early days, researchers used machine learning methods to detect intention. For example, [8] compared the effects of the bag-of-words model, TF-IDF and n-gram methods in short text intention analysis. [9] employ continuous bag-of-words coupled with support vector machines (SVM) to tackle the problem of intention classification.

The wide application of deep learning technology provides a powerful platform for intent detection, and achieves better results [10], such as, [11] presented a novel intention detection system which is based on a self-attention network and a Bi-LSTM. [12] proposed a novel approach to intention detection which involves combining transformer architecture with capsule networks. [13] developed an intention classification model using BERT for the classification of questions received from the users or humans to specific intents regarding the usage of specific features and components of the car. [14] introduce intention detection methods backed by pretrained dual sentence encoders such as USE and ConveRT.

In recent years, multimodal technology has developed rapidly and become a research hotspot in the field of artificial intelligence. It has been widely applied in multiple fields. For example, in emotion recognition [15], multimodal technology can be used to analyze text and image information, identify users' emotional tendencies and expressions. In terms of humor detection [16], various information such as text, speech, and facial expressions are used to determine whether a sentence or situation is

humorous. However, few studies have applied multimodal techniques to intention detection. [5] proposed a model to capture the complex meaning multiplication relationship between image and text in multimodal Instagram posts. [6] proposed a late-fusion approach for the integration of the video signal with the captions signal for intention detection. Other related work is summarized and highlighted in Table 2.1.

Table 2.1: Related work on intention detection

Title	Model/Method	Pros	Cons
Dictionary-based and Rule-based			
Finding suggestions and buy wishes from product reviews [17]	Rules and Graphs	easy to realize. fast operation speed	high workload poor stability
Machine Learning			
intention classification for dialogue utterances [18]	Naive Bayes bag-of-words	considering the order of words	cannot be used in new data
Short Text Intent Classification for Conversational Agent [19]	TF-IDF N-gram	considering the importance of words	inability to accommodate large relation
Intent Detection through Text Mining and Analysis [20]	SVM POS	Part-of-Speech considered	Average performance
Deep Learning			
User Intent Prediction in Information-seeking Conversations [21]	CNN	proving the effectiveness of CNN	CNN can only capture local semantic features
A RNN Contextual Approach to Intent Classification for Goal-oriented Systems [22]	RNN	capturing the features of the entire text	gradient explosion or disappearance for long text
Self-Attention Networks for Intent Detection [23]	Self-attention BILSTM	capturing long-range and multi-scale dependencies	low efficiency
Encoding syntactic knowledge in transformer encoder for intent detection and slot filling [24]	Transformer encoder-based	encode syntactic knowledge into the model	only can be used for text
Intent recognition model based on sequential information and sentence features [25]	CNN BILSTM BERT	leverages contextual and semantic information	require higher computational and storage resources

		within the text	
Multimodal			
An effective multimodal representation and fusion method for multimodal intent recognition [7]	Attention BERT Faster R-CNN	complementarity and consistency are considered	model structure is complex. require higher computational and storage resources

2.2. Multimodal Pre-training

With the gradual maturity of pre-training model technology in the field of natural language, multimodal pre-training models have gradually attracted attention, and a series of visual-language pre-training work has emerged. Vision-and-Language Pre-training VLP (Vision-and-Language Pre-training) [26] refers to a universal representation of cross-modal training based on massive image-text data. The resulting pre-training model can be directly fine-tuned to adapt to downstream vision- language tasks. According to the different encoding methods, it can be roughly divided into twin-tower encoding and fusion encoding.

Twin-tower coding mainly focuses on the representation alignment of the respective modal encoding of images and texts, using the simplest dot product fusion features. Currently hot models such as CLIP [27] and ALIGN [28], etc., this type of method uses contrastive learning for pre-training, uses cosine similarity to measure the distance between modalities, and have demonstrated excellent performance in different fields. [29] proposed a novel visual language pre-training framework ALBEF, which adds an intermediate amount of image-text contrast loss between the image encoder and the text encoder to enable the multimodal encoder to perform better cross-modal alignment. [30] proposed TCL with triple contrastive learning by leveraging cross-modal and intra-modal self-supervision. TCL further considers intra-modal supervision to ensure that the learned representation is also meaningful in each modality, which facilitates cross-modal alignment and joint multi-modal embedding learning. [31] proposed BLIP,

hoping to train a unified multi-modal pre-trained model to solve multi-modal understanding and generation tasks simultaneously. BLIP is a hybrid multi-modal encoder-decoder that can encode images or text in a single mode, image-based text coding and image-based text decoding. Recently, Meta AI He Kaiming's team launched the FILIP [32] multimodal pre-training model, which integrates the image-text double masking technology in MAE [33] and can learn from more image-text data sets in a limited time, and effectively improves the efficiency of model pre-training compared with CLIP.

The fusion coding framework uses the Transformer mechanism for cross-modal fusion. ViLBERT [34] and LXMERT [35] proposed to use three different Transformers for image coding, text coding and feature fusion respectively. After increasing the network depth in the fusion stage, the hybrid coding model framework performed well in visual-language downstream tasks, shows excellent characterization capabilities. However, this type of algorithm is limited by network training and inference speed and has not been widely used in the industry. [36] proposed SimVLM. Different from the general multi-modal pre-trained model using MLM, SimVLM uses the prefixLM method to preserve the visual language representation. [37] proposed ImageBERT, and the authors divided the pre-training process into two parts, first training the model with a large amount of out-of-domain data, and then training with a small amount of in-domain data, so as to get better results on the target task. ViLT [38] is optimized for the inference speed problem. Through a simplified network design, the encoder of the Transformer model is used to extract and process visual features instead of a separate computer vision model to extract features. Experiments show that this method can significantly reduce the number of parameters and running time, and the model effect is significantly better than fusion coding frameworks such as LXMERT, but there is still a certain gap between it and the CLIP twin-tower framework.

After analyzing representative methods in the field of intention detection in recent years, it can be concluded that the research content of multimodal technology mainly includes three parts: feature extraction, multimodal representation, and multimodal

fusion. However, most researchers only focus on the multimodal fusion part and propose some new methods, but the multimodal feature extraction and representation part is less considered and only the traditional single-modal feature extraction method is used. However, the development of multimodal large-scale pre-trained models gives us new ideas for feature extraction and representation.

Chapter-3: Proposed Methodology

This research aims at detecting intention using image and text multimodal data. The main architecture is shown in Figure 3.1, which mainly includes three parts: feature representation, multimodal fusion, and classification. First, in the first part, text and image feature extraction, alignment, and multimodal representation are automatically achieved using the CLIP multimodal large-scale pre-trained model. Secondly, in the second part, considering that different modalities are data of different natures, contain different amounts of information, and have different contributions to intention detection, we design multi-level cross-modality attention module to fuse feature of image-text. Finally, the fused features are input into the classifier to achieve intention detection.

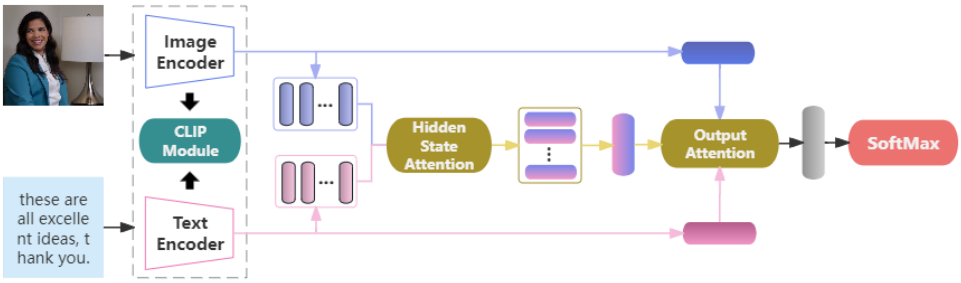

Fig.3.1. Intention detection architecture diagram

3.1. Multimodal Representation

The quality of input features has an important impact on the prediction results of multimodal intention detection models. As early as the machine learning period, feature engineering determined the upper limit of learning. Better features mean you don't need complex models to get excellent results. With the development of deep learning neural networks, the method of feature representation has also changed greatly. Currently, in multimodal intention detection, BERT and ResNet pre-trained models are mainly used to extract text and image features. BERT and ResNet are usually trained independently, this result in each model only understanding the information of its specific modality, which limits the comprehensive understanding of the global context of the model, and the association information between modalities is very critical in image-text tasks. The

multimodal pre-training model has some obvious advantages over the single-modal pre-training model. It uses contrastive learning and other methods to learn the correlation information between modalities in the pre-training stage, so it can process multimodal data at the same time and improve the ability of information understanding. In many image-text tasks, it surpasses the old single-modal scheme and shows strong transfer ability. Moreover, a single multimodal pre-trained model can be directly used to handle multimodal tasks, simplifying the integration and management of the system. Therefore, this study is the first to use the multimodal pre-trained model CLIP in the field of intention detection to extract the features of text and images and achieve multimodal representation.

CLIP (Contrastive Language-Image Pre-training) model is a multimodal pre-training model developed by OpenAI based on 400 million image-text data pairs. It performs well in text and image processing tasks and achieves state of the art performance (SOTA) in many tasks. It uses a contrastive learning method for pre-training, which maps images and text to a common embedding space by maximizing the similarity between relevant image and text pairs while minimizing the similarity between irrelevant image and text pairs, which enables CLIP to understand text and images simultaneously. CLIP is pre-trained on a large-scale multimodal data set. This large-scale data set helps the model learn more general features and can also be fine-tuned on specific tasks to adapt the model to specific fields or applications, thus having versatility and portability, and being able to adapt to different application scenarios. As shown in the figure below, CLIP mainly consists of two parts: Text Encoder and Image Encoder. Text Encoder is used to extract text features and can use the masked self-attention Transformer common in NLP; while Image Encoder is used to extract image features and can adopt the latest proposed ViT-B/16 Transformer architecture.

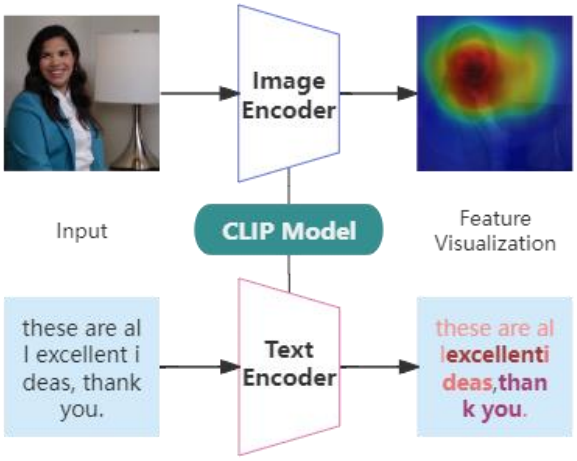

Fig.3.2. Feature representation based on CLIP.

3.1.1. Image Encoder

ViT-B/16 Transformer architecture is used for image coding. It is an image classification model based on Transformer, where ViT represents Vision Transformer, B represents the basic version, and 16 represents that the image is divided into 16×16 image blocks. Compared with traditional convolutional neural networks (CNN), the ViT model adopts a pure Transformer structure, treating images as a series of patch sequences for processing, and has better global perception capabilities and generalization performance. In addition, the ViT model also has the advantage of being highly scalable and can improve performance by increasing the depth and width of the model.

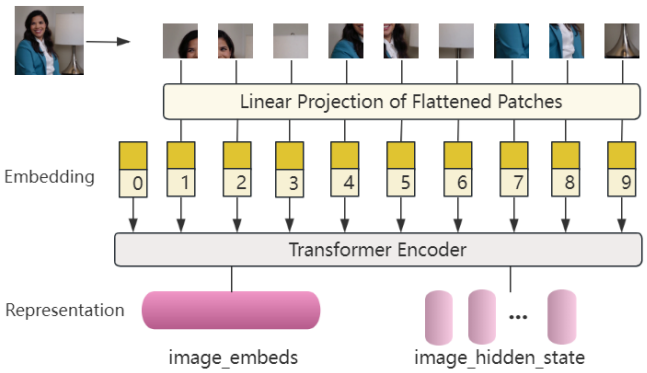

Fig.3.3. Vision Transformer architecture diagram

First, divide the image into patches and linearly transform each patch to obtain input vectors:

$$X = [x_1, x_2, \ldots, x_N] \tag{1}$$

Then, map input vectors xi to embedding vectors zi, add positional information to consider the sequence of inputs:

$$z_i = W_{\text{embed}} \cdot x_i + b_{\text{embed}} + PositionalEmbedding(i) \tag{2}$$

In Transformer Encoder, calculate attention weights and output using self-attention mechanism:

$$\begin{aligned} \text{Attention}(Q, K, V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \\ \text{AttentionOutput} &= \text{Attention}(W_Q \cdot Z, W_K \cdot Z, W_V \cdot Z) \end{aligned} \tag{3}$$

Where, $Q=W_Q \cdot Z$, $K=W_K \cdot Z$, $V=W_V \cdot Z$ are the query, key, and value obtained by linear transformation.

3.1.2. Text Encoder

Masked Self-Attention Transformer is a deep learning method based on Transformer architecture, which is mainly used to process sequence data in text and has strong representation ability and generalization ability. By adopting the Masked Self-Attention mechanism, enables the model to focus on different parts in the input sequence and generate corresponding outputs based on context information.

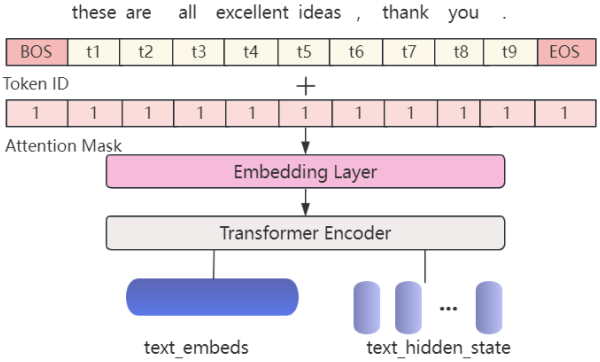

Fig.3.4. Text Transformer architecture diagram.

Tokenize input text into tokens, embedding layer transforms each token t_i into an embedding vector e_i:

$$\begin{aligned} T &= [t_1, t_2, \ldots, t_N] \\ E &= [e_1, e_2, \ldots, e_N] \end{aligned} \tag{4}$$

Apply multiple transformer encoder blocks to process the token embeddings and positional encodings, the encoder structure is the same as in image encoder, then

aggregate the output embeddings of the transformer blocks:

$$\begin{aligned} E_{\text{pos}} &= E + \text{PositionalEncoding}\,(1, 2, \ldots, N) \\ E_{\text{encoded}} &= \text{TransformerEncoder}\,(E_{\text{pos}}) \\ \text{TextEmbedding} &= \text{MeanPooling}\,(E_{\text{encoded}}) \end{aligned} \tag{5}$$

3.2. Multimodal Fusion

In the intention detection task based on image and text fusion, in addition to extracting the features of different modalities, it is more important to fuse the features of different modalities. Multimodal feature fusion is an important process for the model to integrate multiple modalities for prediction tasks. Due to the complementarity and difference between different modal data, the contribution to the results is also different. Feature fusion can provide more effective information for model prediction and improve the accuracy of prediction.

Currently, common multimodal fusion strategies are feature-level fusion [39], decision-level fusion [40] and hybrid fusion [41]. Decision-level fusion can use suitable models for training for different modalities, so it can better extract the internal information of a single modality and has good generalization. However, each modality uses different models for training, which cannot well capture the interaction information between different modalities and is easy to ignore the correlation between different modalities. The hybrid fusion method is flexible in design and has the advantages of both feature-level fusion and decision-level fusion. However, this method is relatively complex, difficult to implement, can easily cause over-fitting problems, and is suitable for scenarios with three modalities and above.

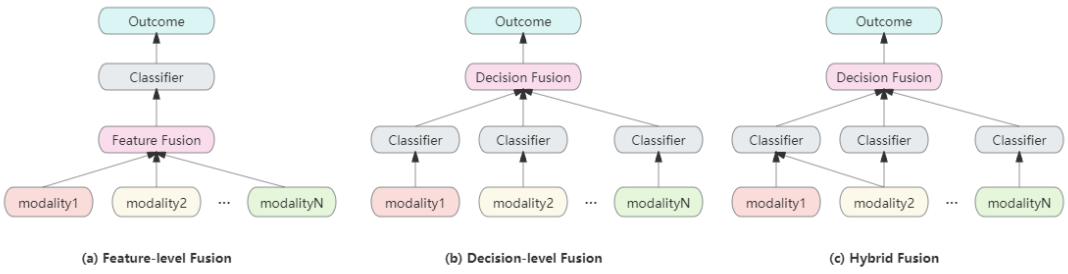

Fig.3.5. Three different fusion strategies

To extract deep features of different modalities and better integrate information between different modalities, this study adopts a feature-level fusion strategy to fuse image and text features based on a cross-modal attention mechanism. Different from the simple vector splicing method, based on Multimodal fusion with cross-modal attention mechanism refers to using the attention mechanism to dynamically adjust the attention between modalities when processing multimodal data to achieve more effective information fusion. In multimodal fusion, the cross-modal attention mechanism allows the model to dynamically adjust the attention to different modalities at each moment, capturing important features while excluding noise. In this way, the model can better understand the overall structure of the multimodal data, thereby improving the performance of the task.

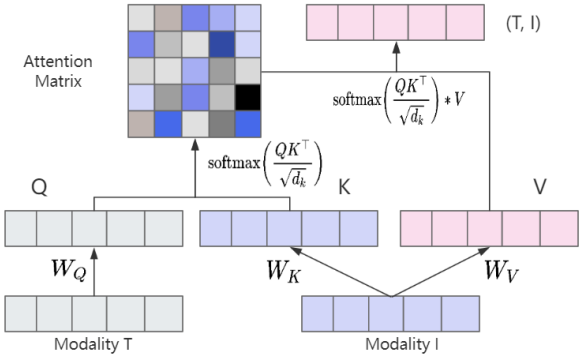

Fig.3.6. Cross-attention calculation process.

We have an image representation $I=[i_1,i_2,\ldots,i_N]$ and a text representation $T=[t_1,t_2,\ldots,t_M]$, where each i and t are feature vectors.

For the image (I) and text (T), calculate Queries (Q), Keys (K), and Values (V):

$$\begin{aligned} \mathbf{Q}_I &= \mathbf{I}\cdot\mathbf{W}_{Q_I}, \quad \mathbf{K}_I = \mathbf{I}\cdot\mathbf{W}_{K_I}, \quad \mathbf{V}_I = \mathbf{I}\cdot\mathbf{W}_{V_I} \\ \mathbf{Q}_T &= \mathbf{T}\cdot\mathbf{W}_{Q_T}, \quad \mathbf{K}_T = \mathbf{T}\cdot\mathbf{W}_{K_T}, \quad \mathbf{V}_T = \mathbf{T}\cdot\mathbf{W}_{V_T} \end{aligned} \tag{6}$$

Text to Image Attention: calculate attention scores for Text Query (Q_T) and Image Key (K_I):

$$\text{Attention}_{T\to I} = \text{softmax}\left(\frac{\mathbf{Q}_T\cdot\mathbf{K}_I^T}{\sqrt{d_k}}\right) \tag{7}$$

Weighted sum of Image Values (V_I) using the attention scores:

$$\text{Output}_{T \to I} = \text{Attention}_{T \to I} \cdot \mathbf{V}_I \tag{8}$$

3.3. Classification

We input the vector obtained by the fusion layer into the multi-layer perceptron. For the intention detection in this article, it is essentially a multi-classification problem. SoftMax can be used as the last layer of the neural network to calculate the intention prediction score. SoftMax is an activation function that normalizes a numeric vector into a probability distribution vector, and the sum of each probability is 1.

$$\text{Softmax}(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \tag{9}$$

$$\hat{y} = \text{softmax}(W * MLP(h) + b) \tag{10}$$

Using Cross Entropy as the loss function, Cross Entropy is an important concept in Shannon information theory and is mainly used to measure the difference in information between two probability distributions.

$$\text{Loss} = -\sum_{i=1}^{n} y_i \log \hat{y}_i \tag{11}$$

n is the total number of intentions, y is the one-hot representation of the sample label, and $\hat{y}$ represents the probability that the sample belongs to the i-th category.

References

[1] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2), 423-443.

[2] Liang, P. P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., ... & Morency, L. P. (2021). Multibench: Multiscale benchmarks for multimodal representation learning. arXiv preprint arXiv:2107.07502.

[3] Obuchowski, A., & Lew, M. (2020, April). Transformer-capsule model for intention detection (student abstract). In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 10, pp. 13885-13886).

[4] Wang, J., Wei, K., Radfar, M., Zhang, W., & Chung, C. (2021, May). Encoding syntactic knowledge in transformer encoder for intention detection and slot filling. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 16, pp. 13943-13951).

[5] Kruk, J., Lubin, J., Sikka, K., Lin, X., Jurafsky, D., & Divakaran, A. (2019). Integrating text and image: Determining multimodal document intention in instagram posts. arXiv preprint arXiv:1904.09073.

[6] Maharana, A., Tran, Q. H., Dernoncourt, F., Yoon, S., Bui, T., Chang, W., & Bansal, M. (2022, July). Multimodal intention Discovery from Livestream Videos. In Findings of the Association for Computational Linguistics: NAACL 2022 (pp. 476-489).

[7] Huang, X., Ma, T., Jia, L., Zhang, Y., Rong, H., & Alnabhan, N. (2023). An Effective Multimodal Representation and Fusion Method for Multimodal intention detection. Neurocomputing, 126373.

[8] Kuchlous, S., & Kadaba, M. (2020, December). Short text intention classification for conversational agents. In 2020 IEEE 17th India Council International Conference (INDICON) (pp. 1-4). IEEE.

[9] Schuurmans, J., & Frasincar, F. (2019). intention classification for dialogue utterances. IEEE Intelligent Systems, 35(1), 82-88.

[10] Louvan, S., & Magnini, B. (2020). Recent neural methods on slot filling and intention classification for task-oriented dialogue systems: A survey. arXiv preprint arXiv:2011.00564.

[11] Yolchuyeva, S., Németh, G., & Gyires-Tóth, B. (2020). Self-attention networks for intention detection. arXiv preprint arXiv:2006.15585.

[12] Obuchowski, A., & Lew, M. (2020, April). Transformer-capsule model for intention detection (student abstract). In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 10, pp. 13885-13886).

[13] Chakraborty, S., Ohm, K. Y., Jeon, H., Kim, D. H., & Jin, H. J. (2023, February). intention Classification of Users Conversation using BERT for Conversational Dialogue System. In 2023 25th International Conference on Advanced Communication Technology (ICACT) (pp. 65-69). IEEE.

[14] Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020). Efficient

intention detection with dual sentence encoders. arXiv preprint arXiv:2003.04807.
[15] Dashtipour, K., Gogate, M., Cambria, E., & Hussain, A. (2021). A novel context-aware multimodal framework for persian sentiment analysis. Neurocomputing, 457, 377-388.
[16] Hasan, M. K., Lee, S., Rahman, W., Zadeh, A., Mihalcea, R., Morency, L. P., & Hoque, E. (2021, May). Humor knowledge enriched transformer for understanding multimodal humor. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 14, pp. 12972-12980).
[17] Ramanand, J., Bhavsar, K., & Pedanekar, N. (2010, June). Wishful thinking-finding suggestions and'buy'wishes from product reviews. In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text (pp. 54-61).
[18] Schuurmans, J., & Frasincar, F. (2019). Intent classification for dialogue utterances. IEEE Intelligent Systems, 35(1), 82-88.
[19] Kuchlous, S., & Kadaba, M. (2020, December). Short text intent classification for conversational agents. In 2020 IEEE 17th India Council International Conference (INDICON) (pp. 1-4). IEEE.
[20] Akulick, S., & Mahmoud, E. S. (2017). Intent detection through text mining and analysis.
[21] Qu, C., Yang, L., Croft, W. B., Zhang, Y., Trippas, J. R., & Qiu, M. (2019, March). User intent prediction in information-seeking conversations. In Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (pp. 25-33).
[22] Mensio, M., Rizzo, G., & Morisio, M. (2018, April). Multi-turn qa: A rnn contextual approach to intent classification for goal-oriented systems. In Companion Proceedings of the The Web Conference 2018 (pp. 1075-1080).
[23] Yolchuyeva, S., Németh, G., & Gyires-Tóth, B. (2020). Self-attention networks for intent detection. arXiv preprint ar**v:2006.15585.
[24] Wang, J., Wei, K., Radfar, M., Zhang, W., & Chung, C. (2021, May). Encoding syntactic knowledge in transformer encoder for intent detection and slot filling. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 16, pp. 13943-13951).
[25] Wu, T., Wang, M., **, Y., & Zhao, Z. (2024). Intent recognition model based on sequential information and sentence features. Neurocomputing, 566, 127054.
[26] Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., & Han, J. (2020). Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12655-12663).
[27] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.
[28] Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., ... & Duerig, T. (2021, July). Scaling up visual and vision-language representation learning with noisy text supervision. In International conference on machine learning (pp. 4904-4916).

PMLR.
[29] Li, J., Selvaraju, R., Gotmare, A., Joty, S., **ong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34, 9694-9705.
[30] Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., ... & Huang, J. (2022). Vision-language pre-training with triple contrastive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15671-15680).
[31] Li, J., Li, D., BLIP:Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning (pp. 12888-12900). PMLR.
[32] Li, Y., Fan, H., Hu, R., Feichtenhofer, C., & He, K. (2023). Scaling language-image pre-training via masking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 23390-23400).
[33] Baade, A., Peng, P., & Harwath, D. (2022). Mae-ast: Masked autoencoding audio spectrogram transformer. arXiv preprint arXiv:2203.16691.
[34] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32.
[35] Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490.
[36] Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., & Cao, Y. (2021). Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904.
[37] Qi, D., Su, L., Song, J., Cui, E., Bharti, T., & Sacheti, A. (2020). Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966.
[38] Kim, W., Son, B., & Kim, I. (2021, July). Vilt: Vision-and-language transformer without convolution or region supervision. In International Conference on Machine Learning (pp. 5583-5594). PMLR.
[39] Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., & Poria, S. (2018). Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-based systems, 161, 124-133.
[40] Liao, C. Y., Chen, R. C., & Tai, S. K. (2018, April). Emotion stress detection using EEG signal and deep learning technologies. In 2018 IEEE International Conference on Applied System Invention (ICASI) (pp. 90-93). IEEE.
[41] Nemati, S., Rohani, R., Basiri, M. E., Abdar, M., Yen, N. Y., & Makarenkov, V. (2019). A hybrid latent space data fusion method for multimodal emotion recognition. IEEE Access, 7, 172948-172964.