**Image and Text Fusion for Intent Prediction**

**Abstract**

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 intent detection. Compared with single-modal data such as images and text, multimodal data can contain more information and can more accurately identify user intentions. In this paper, we propose a new intent detection method, which includes two equally important stages of multimodal representation and fusion, to explore the integration of image and text data to enhance the accuracy of intent detection in multimedia content. The effectiveness of our approach for intent detection based on image and text fusion is proved by comparative experiments with the baseline model on the public multimodal intention dataset.

**Key Words:** intent detection, multimodal technology, CLIP, feature representation, multimodal fusion

# Introduction

As an important research direction in the fields of artificial intelligence and computer science, intent 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 intent 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 intent detection, so that the model can better learn and understand the real intention of users from complex massive data. Researchers have proposed many intent detection methods based on deep learning [1, 2]. These methods can automatically learn relevant patterns from massive data and are more efficient and accurate than manual rules or other methods based on traditional machine learning. The application of these methods promotes the progress of intent detection technology in various fields and provides a more accurate and intelligent user interaction experience for intelligent systems. However, using only a single modality (for example, text modality) for intent 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.

In the era of digital multimedia, the field of intent 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. At present, multimodal technology based on image and text fusion has become a research hotspot in the field of artificial intelligence, and some previous work has explored the application in intent detection [3, 4]. However, most of the existing methods 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.

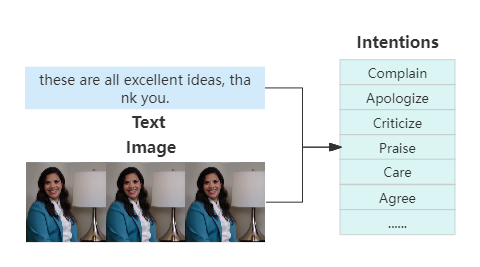


Fig.1. An example of intent detection based on text and image fusion.

In response to these challenges of insufficient single-modal detection capability and the limitations evident in the aforementioned multimodal methods, this research proposes an intent detection method based on image and text fusion. Inspired by the multimodal pre-training model, we use the CLIP model to achieve feature extraction, alignment, and representation. In the fusion stage, we use cross-modal attention mechanism to achieve deep fusion of image-text features, and finally use MLP to output the probability of each intention. The method shows excellent performance on the intent detection dataset compared with baseline models. Our contributions are summarized as follows:

* We propose a multimodal representation method based on multimodal large-scale pre-training model, which uses contrastive learning 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, and it can also simplify system integration and saves computing resources.
* We propose an attention-based Cross-modal multi-level fusion method, which can fuse image and text features from token-level and global level, enables the model to dynamically adjust the attention to different modalities, capture important features, and improve the performance of the model.

# Related Work

## 2.1. Intent detection

In the early days, researchers used machine learning methods to detect intention. For example, [5] compared the effects of the bag-of-words model, TF-IDF and n-gram methods in short text intention analysis. [6] 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 [7], such as, [8] present a novel intent detection system which is based on a self-attention network and a Bi-LSTM [9] propose a novel approach to intent detection which involves combining transformer architecture with capsule networks. [10] 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. [11] introduce intent 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 [12], multimodal technology can be used to analyze text and image information, identify users' emotional tendencies and expressions. In terms of humor detection [13], 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 intent detection. [14] proposed a model to capture the complex meaning multiplication relationship between image and text in multimodal Instagram posts. [3] proposed a late-fusion approach for the integration of the video signal with the captions signal for intent detection. [4] introduced an adaptive multimodal fusion method based on an attention-based gated neural network, which can distinguish the contributions of different modalities.

## 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) [15] 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 [16] and ALIGN [17], 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. [18] 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. [19] 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. [20] 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 FLIP [21] multimodal pre-training model, which integrates the image-text double masking technology in MAE [22] 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 [23] and LXMERT [24] 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. [25] proposed SimVLM. Different from the general multi-modal pre-trained model using MLM, SimVLM uses the prefixLM method to preserve the visual language representation. [26] 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 [27] 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.

# Proposed Methodology

The main architecture is shown in Figure 2, 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 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 intent 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 intent detection.

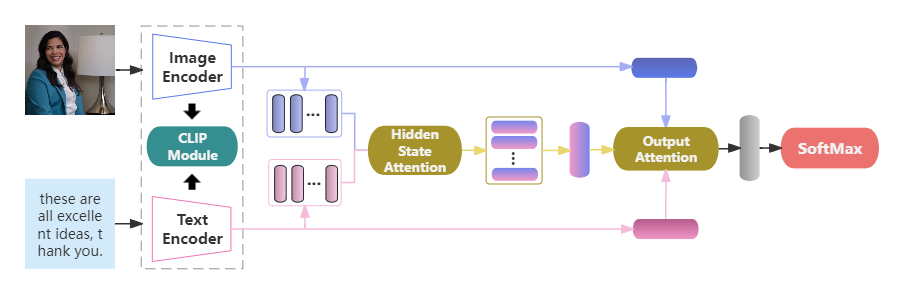


Fig.2. Intent detection architecture diagram

## 3.1. Feature Representation

The quality of input features has an important impact on the prediction results of multimodal intent detection models. As early as the machine learning period, feature engineering determined the upper limit of learning. CLIP (Contrastive Language-Image Pre-training) model is a multimodal pre-training model developed by OpenAI based on 400 million image-text data pairs. 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.

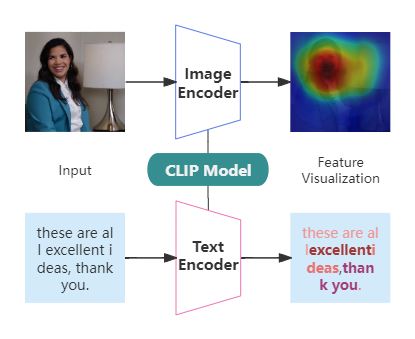


Fig.3. Feature representation based on CLIP.

### 3.1.1. Image Encoder

ViT-B/16 Transformer architecture is used for image coding. It is an image classification model based on Transformer [28], 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.

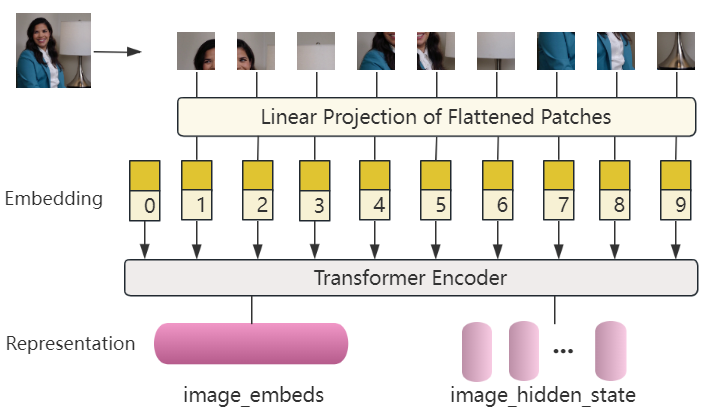


Fig.4. Vision Transformer architecture diagram

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

图片包含 文本

描述已自动生成

(1)

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



(2)

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

文本

描述已自动生成

(3)

Where，*Q*=*WQ*​⋅*Z*，*K*=*WK*​⋅*Z*，*V*=*WV*​⋅*Z* are the query, key, and value obtained by linear transformation.

### 3.1.2. Text Encoder

Masked Self-Attention Transformer is a variant or modification based on Transformer architecture [29], 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.

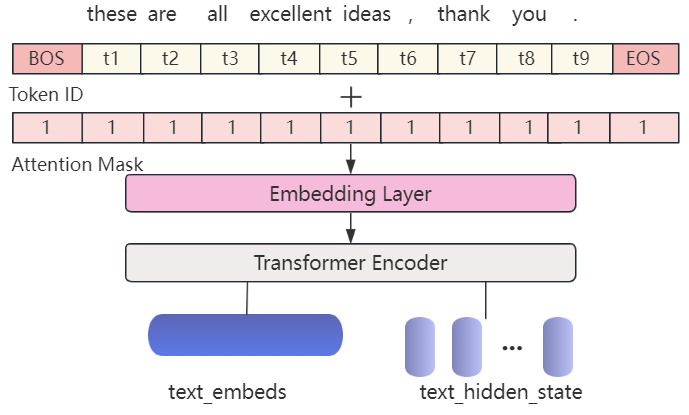
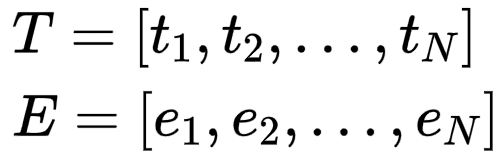


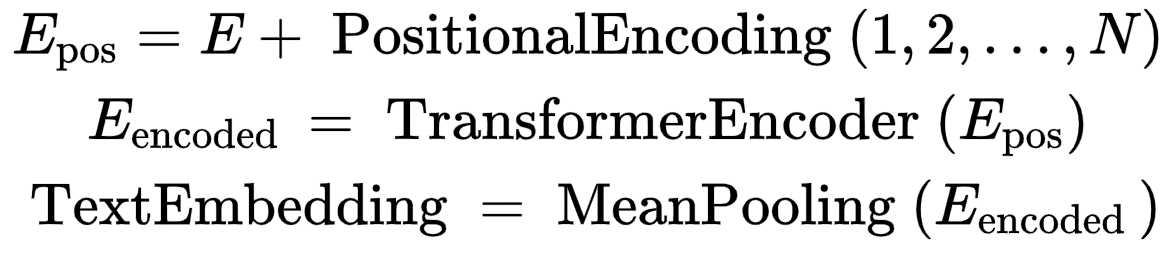
Fig.5. Text Transformer architecture diagram

Tokenize input text into tokens, embedding layer transforms each token *ti*​ into an embedding vector *ei*​:

\

(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:



(5)

## 3.2. Multimodal 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.

To better integrate information between different modalities, this study adopts a multi-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 this way, the model can better understand the overall structure of the multimodal data, thereby improving the performance of the task.

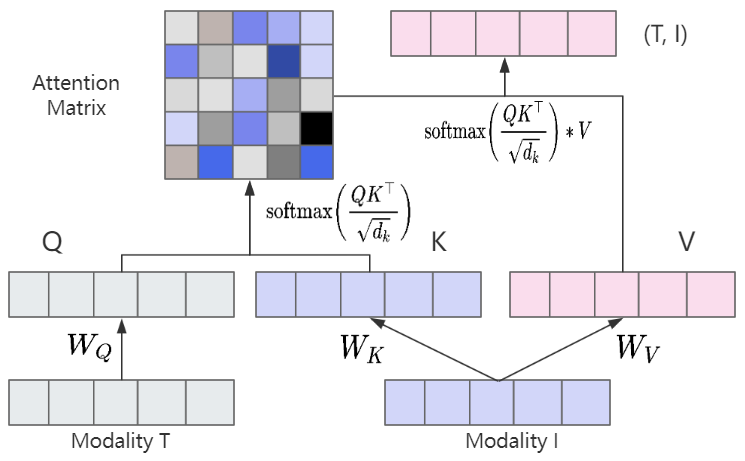


Fig.6. Cross-attention calculation process.

We have an image representation I=[*i*1​,*i*2​,…,*iN*​] and a text representation T=[*t*1​,*t*2​,…,*tM*​], where each i and t are feature vectors.

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

文本

描述已自动生成

(6)

Text to Image Attention: calculate attention scores for Text Query (Q*T*​) and Image Key (K*I*​):

图形用户界面

低可信度描述已自动生成

(7)

Weighted sum of Image Values (V*I*​) using the attention scores:



(8)

## 3.3. Classification

We input the vector obtained by the fusion layer into the multi-layer perceptron. For the intent 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.

形状

中度可信度描述已自动生成

(9)

形状

中度可信度描述已自动生成

(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.

形状

中度可信度描述已自动生成

(11)

n is the total number of intentions, yi is the one-hot representation of the sample label, and i represents the probability that the sample belongs to the i-th category, examples output are as follows:

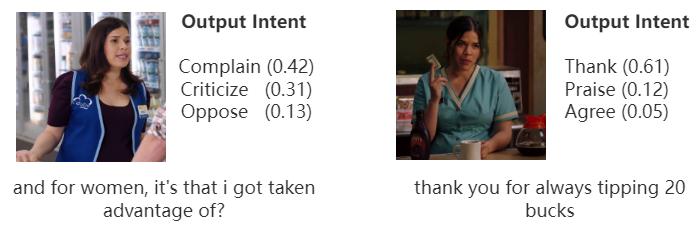


Fig.7. Examples of output, only show top 3 intents.

# Experimental Results

In this part, we train and test the multimodal intent detection method based on image and text fusion proposed in this research, verify the performance of the model on the public dataset through comparative experiments with the baseline model, and complete the ablation experiments of each module of the model.

## 4.1. Dataset and Evaluation

The experiment uses the latest public multimodal intent detection dataset MIntRec [30], which is organized and released by Tsinghua University in 2022. The data comes from the American TV series Superstore, with 2224 high-quality multimodal intention samples screened. Each sample contains three modal information of text, picture, and audio, as well as multimodal intent labels. This dataset combines multimodal scenes to construct a new hierarchical intent system, including two coarse-grained and 20 fine-grained intent categories. The detailed statistics of these datasets are given in Table 1, we split training, validation, and testing sets in 6:2:2. The detailed statistics are shown in Table 2.

Table 1: The statistics of MIntRec.

|  |  |  |
| --- | --- | --- |
| **Coarse-Grained** | **fine-grained** | **Number** |
| Express emotions and attitudes | Complain | 286 |
| Praise | 213 |
| Apologize | 136 |
| Thank | 124 |
| Criticize | 117 |
| Care | 95 |
| Taunt | 62 |
| Agree | 59 |
| Flaunt | 52 |
| Oppose | 51 |
| Joke | 51 |
| Achieve goals | Inform | 284 |
| Advise | 122 |
| Arrange | 110 |
| Introduce | 105 |
| Comfort | 88 |
| Leave | 85 |
| Prevent | 73 |
| Greet | 60 |
| Ask for help | 51 |

Table 2: Dataset splits in MIntRec.

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Express emotions and attitudes** | **Achieve goals** | **Total** |
| Train | 765 | 569 | 1,334 |
| Valid | 240 | 205 | 445 |
| Test | 241 | 204 | 445 |

In this experiment, Accuracy, Precision (P), Recall (R), and F1-score are used as the performance evaluation metrics of the model. Accuracy is the most intuitive metric to measure the accuracy of the model. F1-score is a binary classification metric used to evaluate the performance of the model on imbalanced examples, it can be seen as a weighted average of precision and recall. In the multi-classification problem with imbalanced data samples, we use the macro score for the last three metrics to evaluate the performance of the model.

## 4.2. Implementation Details

In the experiments, we use the Pytorch and HuggingFace Transformers frameworks to develop and train models. In the feature extraction part of the model, clip (clip-vit-base-patch16) is used to extract text and image features simultaneously. In the cross-modal fusion stage, an 8-head cross-attention, 6-layer 512-dimensional Transformer is used. In the classification stage, limited by the size of the dataset, to avoid over-fitting, we use a 2-layer simple neural network as MLP. The dimensions of the last layer are consistent with the number of intention labels, and each value represents the probability of the corresponding label. In the training phase of the model, the pre-trained CLIP weights are used as the initial weights of the image encoder and text encoder in this model, and the weights in the cross-modal attention module and MLP classifier are randomly initialized. Other main hyperparameters are shown in the Table 3. The hyperparameter settings are mainly determined through observation results and based on prior knowledge.

Table 3: Main hyperparameters setting.

|  |  |
| --- | --- |
| **Name** | **Value** |
| Batch Size | 16 |
| Epoch | 15 |
| Learning Rate | 1e-05 |
| Optimizer | Adam |
| Loss Function | Cross Entropy |
| Activation Function | ReLu |
| Dropout Rate | 0.2 |
| Early Stop | 8 |
| Text Dimensions | 512 |
| Image Dimensions | 512 |

## 4.3. Experiments on Intent Detection

To verify the effectiveness of this method proposed in this study, three mainstream multimodal learning models and two mainstream single-modal learning models were selected for comparison with the method:

[31] presentd a novel multimodal learning model using BiGRU and attention-based hybrid fusion strategy (BAHFS). The model uses BiGRU to extract single-modal features and performs pairwise fusion of single-modal features through the dual-modal attention fusion module.

[4] proposed an effective multimodal representation and fusion method (EMRFM), which introduced an adaptive multimodal fusion method based on an attention-based gated neural network, which can distinguish the contributions of different modalities.

[32] proposed a Contrastive Learning and Multi-Layer Fusion Method (CLMLF), which uses contrastive learning strategy to reduce the impact of noise data on the model. In the fusion module, the multi-head attention mechanism is used to realize the alignment and fusion of text and image features.

[33] proposed an intent detection model CBLMA-B (CNN-BiLSTM-Multi-head Attention and BERT), which integrates sequential information and sentence structural features. This approach effectively leverages contextual and semantic information within the text.

[34] proposed Fully Quantized Vision Transformer (FQ-ViT), which utilize a simple yet efficient post-training method that can achieve accurate quantization on Layer Norm inputs with only one-layer wise quantization scale.

Among them, BAHFS, EMRFM, and CLMLF are currently state-of-the-art models of multimodal learning. They are based on the attention mechanism and comprehensively consider the representation, alignment, and fusion of different modal features. CBLMA-B and FQ-ViT are single-modal models, used to process text and images respectively, and are also representative models in the fields of NLP and CV. During the experiment, the parameter settings of the benchmark model mainly referred to the default values, and to ensure the unity of the used modalities, all models only use the picture and text modalities.

Table 4: Results for multimodal intent detection.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Models** | **Modalities** | **ACC** | **F1** | **P** | **R** |
| FQ-ViT | Image | 17.98 | 8.46 | 8.8 | 8.15 |
| CBLMA-B | Text | 69.89 | 67.21 | 67.73 | 66.69 |
| BAHFS | Text + Image | 70.8 | 67.86 | 68.54 | 67.2 |
| CLMLF | Text + Image | 71.46 | 68.13 | 69.32 | 66.98 |
| EMRFM | Text + Image | 71.46 | 68.34 | 69.2 | 67.5 |
| **OURS\*** | Text + Image | **71.91** | **68.59** | **69.44** | **67.77** |

From the comparative experimental results, we can draw the following conclusions. Firstly, from the perspective of overall metrics, the multimodal learning method proposed in this study shows excellent performance on the intent detection dataset compared with other representative baseline models, which verifies the effectiveness of the method. Secondly, from the perspective of input modalities, the results of multimodal models are generally better than the results of single-modal models, because more effective information can be provided with the increase of input modalities, which shows the necessity of fusing multimodal information for intent detection. In addition, in terms of a single modality, the text modality achieved the best performance, which shows that text contains more intent information than images in this dataset, and thanks to the development of large-scale pre-trained language models, text can obtain better semantic representation through transfer learning methods. Using the image modality alone has the worst effect, this may be because the features in the image are scattered and there is a lot of noise, making it difficult for the model to obtain effective features related to the intention from the image.

## 4.4. Ablation Study

To verify the improvement of model performance by each module in this study, ablation experimental studies are carried out on the same dataset for different types of data, feature representation methods, and fusion methods. The experimental results are shown in the Table 5, where "-Text" means removing text data and using empty strings instead, "-Vision" means removing image data and using blank pictures instead, "-CLIP" means removing the CLIP module, Bert and ResNet are used instead to extract text and image features respectively. "-CAF" means remove the cross-attention feature fusion module and use concat method to fuse features.

Table 5: Ablation results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ablation Module** | **ACC** | **F1** | **P** | **R** |
| - Text | 16.63 | 7.65 | 7.86 | 7.45 |
| - Vision | 68.99 | 66.78 | 66.21 | 67.36 |
| - CLIP | 70.11 | 67.14 | 67.09 | 67.19 |
| - CAF | 68.76 | 66.69 | 66.08 | 67.32 |

As can be seen from the first two rows, after removing text, only using image data has the worst effect, with accuracy and F1 score of only 16.63% and 7.65% respectively. This shows that text features play an important role in intent detection, and the role of image information is mainly to extend text information. Intent detection that only relies on visual features is difficult to be put into practical use. In contrast, when only text information data is used for intent detection after removing images, the accuracy is close to 0.7, which is not too far behind the multimodal baseline model in performance, indicating that the text features used in this study are highly relevant to the ideas that users want to express. It can be seen from the third row that the effect decreases after using Bert and ResNet instead of clip model, this may be because the correlation between modalities during feature extraction and representation is not considered. As can be seen from the fourth row, using a simple concatenation to fuse multimodal features, the performance is even lower than the model using only text modality. This means that although the introduction of visual information on the basis of text information makes the model have richer features, it also produces a lot of redundant information or even noise. It is difficult to directly obtain the internal interaction of two modalities by simply relying on the spatial operation of multimodal information for fusion. Therefore, if the information of the additional modalities is not processed properly, it will have a counterproductive effect on the performance of the model.

## 4.5. Influence of Encoders

We further explore the effects of different Encoder on the results. The CLIP multimodal pre-training model includes text and image encoders. The text encoder mainly uses a transformer structure based on the attention mechanism. According to different image encoders, OpenAI provides two major types of pre-training models, namely the ResNet series based on RNN structure and the ViT series based on transformer structure. The experimental results are shown in the Table 6:

Table 6: results of different Encoder.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Encoder** | **ACC** | **F1** | **P** | **R** |
| CLIP-RN50x16 | 70.56 | 67.92 | 67.90 | **67.95** |
| CLIP-ViT32 | 71.46 | 68.25 | 69.18 | 67.34 |
| CLIP-RN50x64 | 71.69 | 68.53 | 69.40 | 67.69 |
| CLIP-ViT16 | **71.91** | **68.59** | **69.44** | 67.77 |

Through experiments, it is found that different encoders will slightly affect model performance, but the overall difference is not obvious. ResNet and ViT series perform similarly because both are mainstream pre-training models in the field of computer vision. Compared with ViT-B-32, the accuracy and F1 value of ViT-B-16 have increased by 0.45 and 0.34 percentage points respectively, and the performance is the best. In general, smaller patches can capture more fine-grained image features, but the actual effect mainly depends on the characteristics of specific tasks and datasets. Due to the small scale and lack of diversity of the dataset, it is difficult to effectively judge the pros and cons of the different encoders.

## 4.6. Error Analysis

We use the confusion matrix to visually show the prediction effect of each intention to further analyze the cases of incorrect prediction in the test data, as shown in Figure 8, where the horizontal axis and vertical axis represent the predicted label and the true label respectively, and the color represents the prediction probability. The diagonal line is that the predicted label is equal to the true label, and the darker color means higher accuracy under this intention.

图表

描述已自动生成

Fig.8. confusion matrix of test results.

As can be seen from the figure above, the model shows excellent performance in most intentions, but there are also obvious differences in the performance of different intentions. Some intentions have relatively fixed expression patterns and specific contents, such as Thank and Greet, so the model shows better performance in them. However, in some complex scenarios, such as Flaunt, Inform, Taunt, and Joke, the model performs generally, which may be because the expressions of these intentions are diversified, and the content is relatively abstract. To reasonably infer the true intention of the speaker, additional modal information such as audio and movement may be required. It can be seen from the confusion matrix that the model is easy to confuse Inform and Arrange, Complain and Oppose. These categories themselves have high similarity, which is easy to cause misjudgment. These problems also show that there is still huge room for improvement in the multimodal intent detection task in complex scenes.

# Conclusion and Future Work

This research mainly explores the application of image-text fusion technology in multimedia intent detection from two different modalities of image and text. Firstly, we design a multimodal learning method based on a multimodal pre-trained model and multi-level cross-modality attention mechanism to achieve more accurate intent detection. Then, the effectiveness of the proposed model is proved by comparative experiments with the baseline model on the same dataset, and the effectiveness of each module is verified by ablation experiments. Finally, we analyzed the specific performance of the model on different intention labels and the possible reasons.

Due to the limitations of data resources and hardware devices, there are still many shortcomings in this study, and there is room for further improvement. In this research, only the text part and the visual part of the MIntRec dataset are used, the audio modality in the video will be added in the subsequent research to ensure the integrity of the data and further improve the accuracy and generalization ability of multimodal intent detection. At the same time, it is common that partial modal data is missing in multimedia, how to solve the problem of missing modes in the input data is of great significance for the practical application of the model.

# References

1. Obuchowski, A., & Lew, M. (2020, April). Transformer-capsule model for intent detection (student abstract). In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 10, pp. 13885-13886).
2. 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).
3. 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).
4. Huang, X., Ma, T., Jia, L., Zhang, Y., Rong, H., & Alnabhan, N. (2023). An Effective Multimodal Representation and Fusion Method for Multimodal intent detection. Neurocomputing, 126373.
5. 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.
6. Schuurmans, J., & Frasincar, F. (2019). intention classification for dialogue utterances. IEEE Intelligent Systems, 35(1), 82-88.
7. 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.
8. Yolchuyeva, S., Németh, G., & Gyires-Tóth, B. (2020). Self-attention networks for intent detection. arXiv preprint arXiv:2006.15585.
9. Obuchowski, A., & Lew, M. (2020, April). Transformer-capsule model for intent detection (student abstract). In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 10, pp. 13885-13886).
10. 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.
11. Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020). Efficient intent detection with dual sentence encoders. arXiv preprint arXiv:2003.04807.
12. Dashtipour, K., Gogate, M., Cambria, E., & Hussain, A. (2021). A novel context-aware multimodal framework for persian sentiment analysis. Neurocomputing, 457, 377-388.
13. 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).
14. 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.
15. 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).
16. 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.
17. 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.

[18] 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.

[19] 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).

[20] 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.

[21] 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).

[22] Baade, A., Peng, P., & Harwath, D. (2022). Mae-ast: Masked autoencoding audio spectrogram transformer. arXiv preprint arXiv:2203.16691.

[23] 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.

[24] Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490.

[25] 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.

[26] 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.

[27] 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.

[28] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[30] Zhang, H., Xu, H., Wang, X., Zhou, Q., Zhao, S., & Teng, J. (2022, October). Mintrec: A new dataset for multimodal intent recognition. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 1688-1697).

[31] Liu, Z., Zhou, B., Meng, L., & Huang, G. (2023). Multimodal Sentiment Analysis Using BiGRU and Attention-Based Hybrid Fusion Strategy. Intelligent Automation & Soft Computing, 37(2).

[32] Li, Z., Xu, B., Zhu, C., & Zhao, T. (2022). CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection. arxiv preprint arxiv:2204.05515.

[33] Wu, T., Wang, M., \*\*, Y., & Zhao, Z. (2024). Intent recognition model based on sequential information and sentence features. Neurocomputing, 566, 127054.

[34] Lin, Y., Zhang, T., Sun, P., Li, Z., & Zhou, S. (2021). Fq-vit: Post-training quantization for fully quantized vision transformer. ar\*\*v preprint ar\*\*v:2111.13824.