paperv1

by meng wang

Submission date: 20-Dec-2023 02:08AM (UTC-0800)

Submission ID: 2263069544

File name: paper-v1.docx (681.28K)

Word count: 5702

Character count: 32343

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. In this paper, we design a new intention detection framework decomposing the multimodal learning problem into two equally important stages of representation and fusion, to explore the integration of image and text data to enhance the accuracy and robustness of intention detection in multimedia content. The effectiveness of our approach for intention detection based on image and text fusion is proved by comparative experiments with the baseline model on the public multimodal intention dataset.

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. Researchers have proposed many intention detection methods based on deep learning (Obuchowski et al., 2020; Wang et al., 2021). 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 intention 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 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.

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. Recently, multimodal machine learning has become one of the hot research areas, which aims to build models that can process and correlate information from multiple modalities. Some previous work has explored the application of multimodal technology in intention detection. (Maharana et al., 2022) introduces a late-fusion approach for integration of the video signal with the captions signal for intention Detection and shows significant improvements with unimodal pre-trained models. (Huang et al., 2023) develops an adaptive multimodal fusion method based on an attention-based gated neural network in intention detection, which can distinguish the contributions of different modalities. At present, there are few studies on intention detection based on image and text fusion, and 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.

In response to these challenges of insufficient single-modal detection capability and the limitations evident in the aforementioned multimodal methods, this research proposes an intention 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 the attention mechanism to achieve deep fusion of image-text features, and finally use the MLP to predict the intention. The method shows excellent performance on the intent detection dataset compared with baseline models. The main contributions of this paper are 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.

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, (Kuchlous et al., 2020) compared the effects of the bag-of-words model, TF-IDF and n-gram methods in short text intention analysis. (Schuurmans et al., 2019) employ continuous bag-of-words coupled with support vector machines (SVM) to tackle the problem of intention classification.

With the development of deep learning, many intention detection methods based on deep learning have been proposed (Louvan et al., 2020), such as, (Yolchuyeva et al., 2020) present a novel intention detection system which is based on a self-attention network and a Bi-LSTM (Obuchowski et al., 2020) propose a novel approach to intention detection which involves combining transformer architecture with capsule networks. (Chakraborty et al., 2023) 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. (Casanueva et al., 2020) 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 (Dashtipour et al., 2021), multimodal technology can be used to analyze text and image information, identify users' emotional tendencies and expressions. In terms of humor detection (Hasan et al., 2021), 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. (Kruk et al., 2019) proposed a model to capture the complex meaning multiplication relationship between image and text in multimodal Instagram posts. (Maharana et al., 2022) proposed a late-fusion approach for the integration of the video signal with the captions signal for intention detection. (Huang et al., 2023) introduced an adaptive multimodal fusion method based on an attention-based gated neural network, which can distinguish the contributions of different modalities.

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) (Chen et al., 2020) 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 (Radford et al., 2021) and ALIGN (Jia et al., 2021), 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. Recently, Meta AI He Kaiming's team launched the FLIP (Li et al., 2023) multimodal pre-training model, which integrates the image-text double masking technology in MAE (Baade et al., 2022) 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 (Lu et al., 2019) and LXMERT (Tan et al., 2019) 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. ViLT (Kim et al., 2021) 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.

This research aims at detecting intention using image and text multimodal data. 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 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, multimodal feature fusion method based on cross-modal attention mechanism is designed. Finally, the fused features are input into the classifier to achieve intention detection.

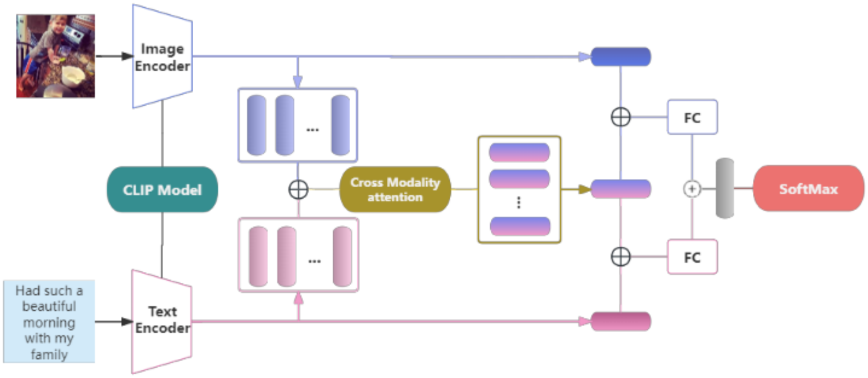


Fig.2. Intention detection architecture diagram

3.1. Feature 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.

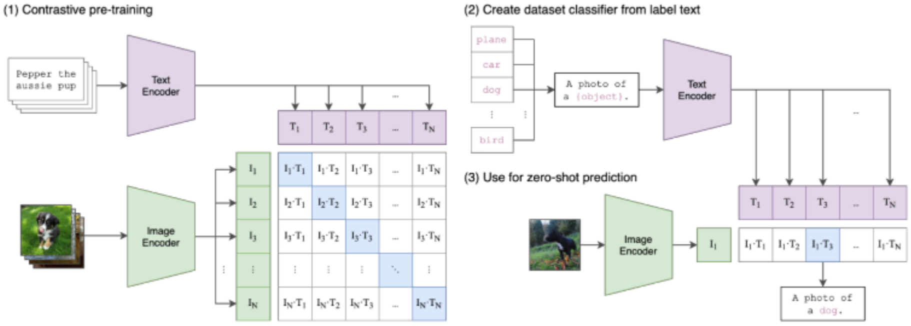

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ... IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Fig.3. CLIP architecture diagram

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.

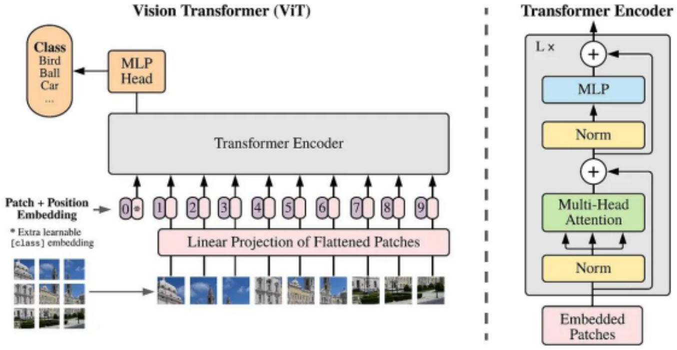


Fig.4. Vision Transformer architecture diagram

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.

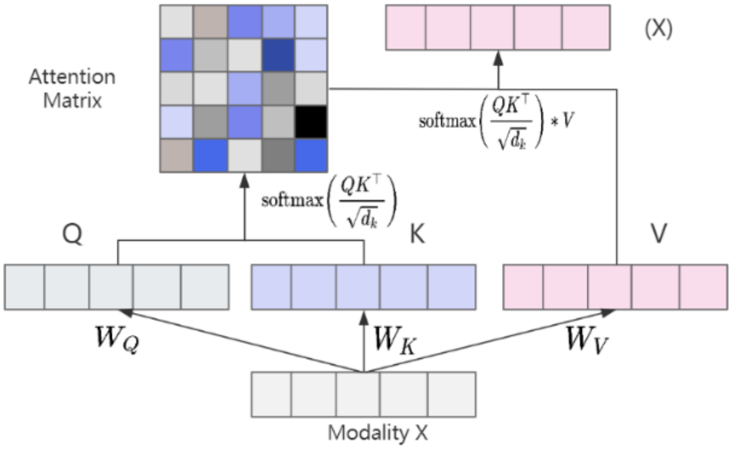
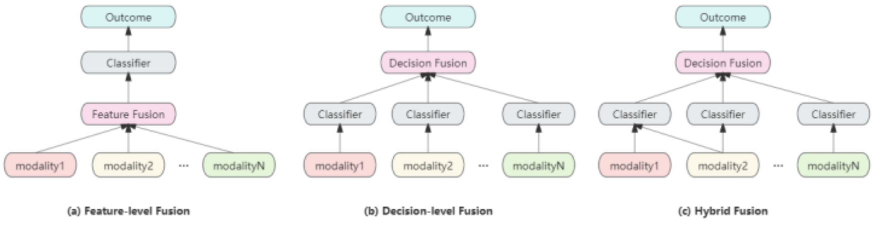

(X)
Attention Matrix
$\mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right) * V$
$\mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)$
Q
K
V
W_Q
W_K
W_V
Modality X

Fig.5. Self-attention calculation process.

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 (Majumder et al., 2018), decision-level fusion (Liao et al., 2018) and hybrid fusion (Nemati et al., 2019). 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.



(a) Feature-level Fusion

(b) Decision-level Fusion

(c) Hybrid Fusion

Fig.6. Three different fusion strategies

In order 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.

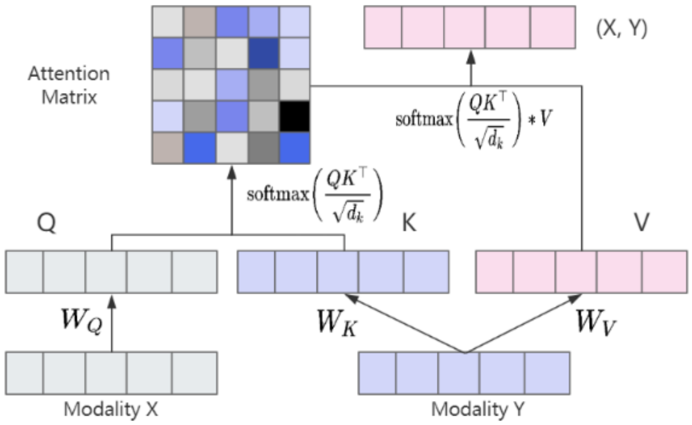

Attention Matrix
(X, Y)
$\mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) * V$
$\mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)$
Q
K
V
W_Q
W_K
W_V
Modality X
Modality Y

Fig.7. Cross-attention calculation process.

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.

$$\mathrm{Softmax}(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \tag{1}$$

$$\hat{y} = \mathrm{softmax}(W * MLP(h) + b) \tag{2}$$

where W and b represent linear layer parameters and bias terms respectively. 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.

$$\text{Loss } = -\sum_{i=1}^{n} y_i \log \hat{y}_i \tag{3}$$

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

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

4.1. Dataset and Evaluation

4.1.1. Dataset

The experiment used the latest public multimodal intent detection dataset (MIntRec) (Zhang et al., 2022), organized and released by Tsinghua University in 2022. MIntRec is a multimodal intent detection dataset, that is mainly used for intent detection in real multimodal scenes and is currently the first benchmark dataset for intent detection in real-world multimodal scenes. 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. Inspired by human intention philosophy and goal-oriented intentions in artificial intelligence research, the data is categorized into two coarse-grained intent categories: "Express emotions or attitudes" and "Achieve goals". "Express emotions and attitudes" contains 11 fine-grained intention categories: Complain, Praise, Apologize, Thank, Criticize, Care, Agree, Taunt, Flaunt, Oppose and Joke. "Achieve goals" are classified into nine categories: Inform, Advise, Arrange, Introduce, Comfort, Leave, Prevent, Greet, and Ask for help. The 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.

First Level	Second Level	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

4.1.2. Evaluation Metrics

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 indicator 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, Micro-F1 or Macro-F1 metrics are usually used to evaluate the performance of the model. We use the macro score over all classes for the last three metrics. The higher values indicate better performance of all metrics.

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 which the image encoder and text encoder use ViT-B/16 and the transformer structure based on the self-attention mechanism respectively. The ViT-B/16 model uses a patch size of 16x16 pixels to extract image features, which means that the input image is divided into 16x16 non-overlapping patches. Each patch is flattened into a 2D vector and fed into the transformer encoder. The number of patches is then reduced by a factor of 96 to obtain a sequence of image features. The ViT-B/16 model uses a patch size of 16x16 pixels to extract image features, which means that the input image is divided into 16x16 non-overlapping patches. Each patch is flattened into a 2D vector and fed into the transformer encoder. The number of patches is then reduced by a factor of 96 to obtain a sequence of image features.

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, in order to avoid over-fitting, a 2-layer MLP, and a SoftMax layer simple classification network were constructed. The dimensions of the SoftMax 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:

MulT (Tsai et al., 2019). The Multimodal Transformer (MulT) is an end-to-end method to address the challenge of processing and understanding information from multiple modalities that may not be temporally synchronized or aligned, MulT extends the Transformer architecture to capture the adaptation knowledge between different modalities in the latent space.

(Rahman et al., 2019) proposed a Multimodal Adaptation Gate architecture (MAG), which is an improved version of BERT-based models that allows the model to input non-textual modalities. It can be flexibly placed between layers of BERT. The input of different modalities will affect the meaning of the words, which in turn affects the position of the vector in the semantic space, and MAG can produce a position shift to recalculate the new position of the vector in the semantic space.

Trans_TAV. This model is a relatively simple multimodal learning method, which utilizes an early fusion approach for combining features from different modalities. The method can use BERT to extract text information, and Wav2vec and Faster R-CNN to extract audio and video information respectively.

(Kenton et al., 2019) BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained natural language processing (NLP) model that adopts the Transformer architecture and is pre-trained on a large-scale text corpus to learn universal language representations.

(He et al., 2016) ResNet-50 is a pre-trained model for images, mainly used for image classification tasks. It is also often used as a basic model for transfer learning to handle various computer vision tasks.

Among them, MulT, MAG-BERT, and Trans_TAV are representative models of multimodal learning. The first two are based on the attention mechanism and comprehensively consider the representation, alignment, and fusion of different modal features. Compared with Trans_TAV, they are more complex and advanced and have better multimodal learning capabilities. While Trans_TAV is relatively simple to implement but has shortcomings in feature fusion. It is a typical representative of early traditional multimodal learning methods. BERT and ResNet-50 are single-modal models, used to process text and images respectively, and are also representative models in the fields of NLP and CV. Through the comparison with the above five representative models, we can effectively evaluate the performance of the multimodal learning method based on the multimodal pre-training model and cross-modal attention mechanism proposed in this study on intent detection. During the experiment, the parameter settings of the benchmark model mainly referred to the default values, and in order to ensure the unity of the used modalities, all models only use the picture and text modalities.

Table 4: Overall results for multimodal intent detection on the MIntRec dataset.

Methods	Modalities	ACC	F1	P	R
ResNet-50	Image	17.30	7.98	8.10	7.87
Trans_TAV	Text + Image	69.44	67.06	66.70	67.43
BERT	Text	69.89	67.20	67.16	67.25
MulT	Text + Image	71.24	67.85	68.32	67.39
MAG-BERT	Text + Image	71.69	68.59	69.36	**67.83**
OURS*	Text + Image	**71.91**	**68.59**	**69.44**	67.77

Table 4 shows the overall comparative experimental results. From the 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 detection 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. Finally, from the perspective of multimodal models, the Trans_TAV model has the worst effect. This may be because it is difficult to effectively utilize the complementarity between multimodal modes by directly splicing features together or simply using a simple weighted summation method to fuse single-modal features. This also shows that in multimodal learning, it is necessary to design a reasonable multimodal fusion method to effectively utilize multimodal information and thereby improve the performance of the model.

4.4. Ablation Study

In order 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 on the MIntRec dataset.

Model	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
OURS*	**71.91**	**68.59**	**69.44**	**67.77**

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 multimodal learning method is like the Trans_TAV model, which does not consider the correlation between modalities during feature extraction and representation, and it is difficult to accurately fuse the information expressed by different inputs in the subsequent stage. 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. ResNet mainly includes RN50x16 and RN50x64, x16 and x64 mean a scaling factor applied to the number of channels (or filters) in each layer. ViT mainly includes ViT-B/32 and ViT-B/16, 32 and 16 Refer to the patch size used in the input images. In order to verify the impact of different image encoders on performance, the above four encoders were used for comparative experiments. , the experimental results are shown in the Table 6:

Table 6: results of different Encoder on the MIntRec dataset.

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 was 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. This is due to the impact of patch size on model performance. 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.

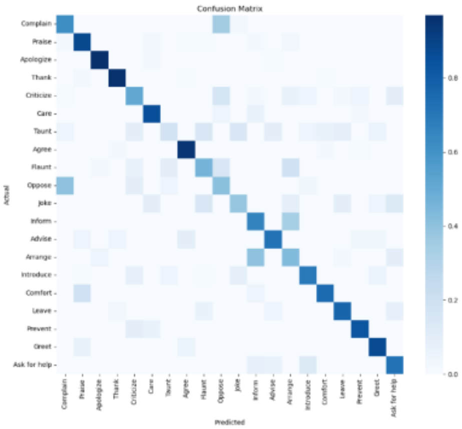


Fig.8. confusion matrix of test results.

Overall, the model shows high accuracy in most categories, but there are also obvious differences in the performance of different intentions. Some intentions have relatively fixed expression patterns and specific contents, such as Praise, Thank, Apologize, Agree, and Greet, and the model shows better performance in these categories. 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.

Chapter-5: Conclusion and Future Work

This research mainly explores the application of image-text information fusion technology in multimedia intent detection from two different modalities of image and text. Firstly, this study designs a multimodal learning method based on a multimodal pre-trained model and a cross-attention mechanism to achieve more accurate intent detection. In order to better utilize the information of these two different modalities, the intent detection task is divided into two parts: multimodal feature representation and fusion. In the feature representation part, we propose to use CLIP multimodal pre-training to extract text and image features simultaneously, and automatically achieve alignment after fine-tuning, which endows the model with the ability to learn with a small number of samples. In the fusion part, the cross-attention mechanism is used to fuse the information of different models, so as to effectively use the interaction information of different modalities and improve the performance of the model. Then, the effectiveness of the proposed model is proved by comparative experiments with the baseline model on the same dataset. Then, 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.

paperv1

ORIGINALITY REPORT

21%	10%	19%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Xuejian Huang, Tinghuai Ma, Li Jia, Yuanjian Zhang, Huan Rong, Najla Alnabhan. "An Effective Multimodal Representation and Fusion Method for Multimodal Intent Recognition", Neurocomputing, 2023 Publication	5%
2	arxiv.org Internet Source	2%
3	www.ncbi.nlm.nih.gov Internet Source	1%
4	www.mdpi.com Internet Source	1%
5	Submitted to Liverpool John Moores University Student Paper	1%
6	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2019 Publication	1%
7	www.arxiv-vanity.com Internet Source	

		1%
8	icact.org Internet Source	1%
9	medium.com Internet Source	1%
10	Submitted to Hong Kong Baptist University Student Paper	1%
11	journals.plos.org Internet Source	<1%
12	"Web Information Systems and Applications", Springer Science and Business Media LLC, 2023 Publication	<1%
13	Submitted to University of Central Lancashire Student Paper	<1%
14	C.R. Dhivyaa, K. Nithya, G. Sendooran, R. Sudhakar, K.Sathis Kumar, S.Santhosh Kumar. "XLNet Transfer Learning Model for Sentimental Analysis", 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), 2023 Publication	<1%
15	preview.aclanthology.org Internet Source	<1%

16	pubs.rsc.org Internet Source	<1 %
17	Kezhong Wang, Ting Jin. "Chapter 3 Multimodal Social Media Sentiment Analysis Based on Cross-Modal Hierarchical Attention Fusion", Springer Science and Business Media LLC, 2022 Publication	<1 %
18	Muhammad Shahid Jabbar, Jitae Shin, Jun-Dong Cho. "AI Ekphrasis: Multi-Modal Learning with Foundation Models for Fine-Grained Poetry Retrieval", Electronics, 2022 Publication	<1 %
19	Xu Zhang, Yanzheng Xiang, Zejie Liu, Xiaoyu Hu, Deyu Zhou. "I2R: Intra and inter-modal representation learning for code search", Intelligent Data Analysis, 2023 Publication	<1 %
20	Zeyu Yu, Hui Fang, Qiannan Zhangjin, Chunxiao Mi, Xuping Feng, Yong He. "Hyperspectral imaging technology combined with deep learning for hybrid okra seed identification", Biosystems Engineering, 2021 Publication	<1 %
21	aclanthology.lst.uni-saarland.de Internet Source	<1 %

22	Tao Hu, Xuyu Xiang, Jiaohua Qin, Yun Tan. "Audio-Text Retrieval Based on Contrastive Learning and Collaborative Attention Mechanism", Research Square Platform LLC, 2022 Publication	<1%
23	Submitted to University of Southern California Student Paper	<1%
24	"Advances in Multimedia Information Processing – PCM 2018", Springer Science and Business Media LLC, 2018 Publication	<1%
25	Chunlun Xiao, Anqi Zhu, Chunmei Xia, Yuanlin Liu et al. "Dual-Branch Multimodal Fusion Network for Skin Lesions Diagnosis using Clinical and Ultrasound Image", 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), 2023 Publication	<1%
26	Song Yang, Qiang Li, Wenhui Li, Xuanya Li, An-An Liu. "Dual-Level Representation Enhancement on Characteristic and Context for Image-Text Retrieval", IEEE Transactions on Circuits and Systems for Video Technology, 2022 Publication	<1%

27	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2020 Publication	<1 %
28	Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, Jiayan Teng. "MIntRec: A New Dataset for Multimodal Intent Recognition", Proceedings of the 30th ACM International Conference on Multimedia, 2022 Publication	<1 %
29	link.springer.com Internet Source	<1 %
30	"Computer Vision – ECCV 2022", Springer Science and Business Media LLC, 2022 Publication	<1 %
31	"TSCA on Students' Attitude toward Writing in Blended Learning of College English", Frontiers in Educational Research, 2023 Publication	<1 %
32	Yi Liu, Yue Zhang, Haidong Hu, Xiaodong Liu, Lun Zhang, Ruijun Liu. "An Extended Text Combination Classification Model for Short Video Based on Albert", Journal of Sensors, 2021 Publication	<1 %
33	artemis-ia.eu Internet Source	<1 %

34	Changning Li, Haiyun Li, Tuo Yao, Ming Su, Fu Ran, Jianhong Li, Li He, Xin Chen, Chen Zhang, Huizhen Qiu. "Effects of swine manure composting by microbial inoculation: Heavy metal fractions, humic substances, and bacterial community metabolism", Journal of Hazardous Materials, 2021 Publication	<1 %
35	Fuqiang Li, Tongzhuang Zhang, Yong Liu, Feiqi Long. "Deep Residual Vector Encoding for Vein Recognition", Electronics, 2022 Publication	<1 %
36	Lei Shi, Shijie Geng, Kai Shuang, Chiori Hori, Songxiang Liu, Peng Gao, Sen Su. "Multi-Layer Content Interaction Through Quaternion Product for Visual Question Answering", ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020 Publication	<1 %
37	Ruiyang Jia, Wei Sun, Zixin Song, Jianjun Wu, Dajian Li. "Chapter 255 A UAV Vision Autonomous Localization Algorithm Based on Multi-indicator Regression of Global Features", Springer Science and Business Media LLC, 2023 Publication	<1 %
	downloads.hindawi.com	

38	Internet Source	<1 %
39	Cong Jin, Armagan Elibol, Pengfei Zhu, Nak Young Chong. "Semantic Mapping Based on Image Feature Fusion in Indoor Environments", 2021 21st International Conference on Control, Automation and Systems (ICCAS), 2021 Publication	<1 %
40	He Wang, Xinshan Zhu, Peiyin Chen, Yuxuan Yang, Chao Ma, Zhongke Gao. "A gradient-based automatic optimization CNN framework for EEG state recognition", Journal of Neural Engineering, 2022 Publication	<1 %
41	Shanshan Du, Yingwen Wang, Xinyu Huang, Rui-Wei Zhao, Xiaobo Zhang, Rui Feng, Quanli Shen, Jianqiu Zhang. "Chest X-ray Quality Assessment Method with Medical Domain Knowledge Fusion", IEEE Access, 2023 Publication	<1 %
42	Zihan Zheng, Ningxia Chen, Jianhao Wu, Zhixuan Xv, Shuangyin Liu, Zhijie Luo. "EW-YOLOv7: A Lightweight and Effective Detection Model for Small Defects in Electrowetting Display", Processes, 2023 Publication	<1 %

43	Hui Li, Xiao-Jun Wu. "CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach", Information Fusion, 2024 Publication	<1 %
44	Shaohong Zhou, Junjiang He, Tao Li, Xiaolong Lan, Yunpeng Wang, Hui Zhao, Yihong Li. "Automating the Deployment of Cyber Range with OpenStack", The Computer Journal, 2023 Publication	<1 %
45	Yuhang Xu, Yingjie Wei, Yangyang Sha, Cong Wang. "A novel model with an improved loss function to predict the velocity field from the pressure on the surface of the hydrofoil", Ocean Engineering, 2023 Publication	<1 %
46	Yuhua Zhu, Hang Li, Tong Zhen, Zhihui Li. "Integrating Self-Attention Mechanisms and ResNet for Grain Storage Ventilation Decision Making: A Study", Applied Sciences, 2023 Publication	<1 %
47	export.arxiv.org Internet Source	<1 %
48	openreview.net Internet Source	<1 %
49	zaguan.unizar.es Internet Source	<1 %

50	"Computer Vision Systems", Springer Science and Business Media LLC, 2023 Publication	<1%
51	Tianqi Zhao, Ming Kong, Tian Liang, Qiang Zhu, Kun Kuang, Fei Wu. "CLAP: Contrastive Language-Audio Pre-training Model for Multi-modal Sentiment Analysis", Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, 2023 Publication	<1%
52	Song Yang, Qiang Li, Wenhui Li, Xuanya Li, Ran Jin, Bo Lv, Rui Wang, An-An Liu. "Semantic Completion and Filtration for Image-Text Retrieval", ACM Transactions on Multimedia Computing, Communications, and Applications, 2022 Publication	<1%
53	Yuxuan Wu, Zhizhong Liu, Zhaohui Su, Xiaoyu Song. "Chapter 13 Multimodal Intent Recognition Based onContrastive Learning", Springer Science and Business Media LLC, 2023 Publication	<1%
54	Zhenze Yang, Markus J. Buehler. "Words to Matter: De novo Architected Materials Design Using Transformer Neural Networks", Frontiers in Materials, 2021 Publication	<1%

55 Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, Jing Shao. "CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval", 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019 <1%

Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		