# Chapter-4: Experimental Section

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, and its calculation method is shown in Formula 4.

形状

中度可信度描述已自动生成

(4)

Among them, TP represents the number of samples whose predicted label is positive and the actual label is also positive, TN represents the number of samples whose predicted label is negative and the actual label is also negative, FP represents the number of samples whose predicted label is positive but the actual label is negative, and FN Represents the number of samples whose predicted label is negative but the actual label is positive.

Similar to accuracy, precision is used to calculate the ratio of true positive predictions to the total number of positive predictions made by the model. The calculation formula is as follows:

形状

中度可信度描述已自动生成

(5)

In addition, Recall represents the proportion of positive samples correctly predicted by the model in all positive samples, and recall and precision are a pair of contradictory metrics. When recall is high, precision is generally low. When precision is high, recall is generally low. It is calculated as follows:

形状

中度可信度描述已自动生成

(6)

However, for datasets with an unbalanced number of positive and negative samples, the above evaluation metrics have certain flaws, because even if the classifier predicts all samples as categories with a larger number, it can still obtain higher accuracy and precision rates, but such a classifier actually has no effect. For data with an unbalanced number of positive and negative samples, a more reasonable evaluation index is the F1 score, and its calculation formula is:

形状

中度可信度描述已自动生成

(7)

Where P is Precision and R is Recall.

F1-score is a binary classification metric used to evaluate the performance of the model on imbalanced examples. The F1-score 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

All the experiments in this part are run on a server equipped with two Tesla T4 GPUs. The server memory is 56GB and the total GPU memory is 32GB. The system is Ubuntu. The experimental code running environment uses Docker image configuration, and the programming language is Python 3.8, the deep learning framework uses PyTorch 2.0.

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 | 1.00E-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 CLIF\_CMA 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 Encoder

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.

图表

描述已自动生成

Fig.8. confusion matrix of test results.

On the whole, 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, gratitude, apology, agreement, and greeting, and the model shows better performance in these categories. However, in some complex scenarios, such as showing off, informing, mocking, and joking, 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 informing and arranging, complaining, and opposing. 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.