**Image and Text Fusion for Intent Detection in Multimedia**

**Wang Meng**

**Abstract**

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

Compared with single-modal data such as images and text, Multimodal data can offer richer information for more precise user intention detection. Currently, there is limited research on intention detection through image and text fusion. Existing studies focus primarily on integrating modal features during feature fusion, utilizing only single-modality pre-trained models for feature extraction. However, this method complicates the system, raises computational costs, and may limit the model's contextual understanding, thus hindering effective capture of inter-modal correlation information.

This research endeavors to design a novel intention detection framework to enhance the accuracy and robustness of intention detection. The proposed method includes two equally important stages of multimodal representation and fusion. In the representation part, we introduce multimodal large-scale pre-training models to extract text and image features simultaneously. In the fusion part, we use an attention-based Cross-modal multi-level fusion method, which can fuse image and text features from token level and global level. To verify the effectiveness of the proposed method, this study develops an intention detection framework and carry out experimental verification on the intention detection dataset. The experimental results show that the model Acc by 0.45-1.1% and the F1 value by 0.25-0.73% over other baseline multimodal models, which verifies the effectiveness of the model in this research. We also use ablation experiments to verify the validity of each module of the model.

**Key Words:** intention detection, multimodal technology, Contrastive Language-Image Pre-Training (CLIP), feature representation, multimodal fusion

**Table of Contents**

[Chapter-1: Introduction 5](#_Toc163503501)

[1.1. Background 5](#_Toc163503502)

[1.2. Motivation 6](#_Toc163503503)

[1.3. Problem Statement 7](#_Toc163503504)

[1.4. Research Significance 9](#_Toc163503505)

[1.5. Research Questions 9](#_Toc163503506)

[1.6. Research Objectives 9](#_Toc163503507)

[1.7. Scope of the Study 10](#_Toc163503508)

[Chapter-2: Literature Review 11](#_Toc163503509)

[2.1. Intention Detection 11](#_Toc163503510)

[2.2. Multimodal Pre-training 13](#_Toc163503511)

[Chapter-3: Proposed Methodology 16](#_Toc163503512)

[3.1. Multimodal Representation 16](#_Toc163503513)

[3.1.1. Image Encoder 18](#_Toc163503514)

[3.1.2. Text Encoder 19](#_Toc163503515)

[3.2. Multimodal Fusion 20](#_Toc163503516)

[3.3. Classification 22](#_Toc163503517)

[Chapter-4: Experimental Section 23](#_Toc163503518)

[4.1. Dataset and Evaluation 23](#_Toc163503519)

[4.1.1. Dataset 23](#_Toc163503520)

[4.1.2. Evaluation Metrics 24](#_Toc163503521)

[4.2. Implementation Details 26](#_Toc163503522)

[4.3. Experiments on Intent Detection 27](#_Toc163503523)

[4.4. Ablation Study 30](#_Toc163503524)

[4.5. Influence of Encoders 31](#_Toc163503525)

[4.6. Error Analysis 32](#_Toc163503526)

[Chapter-5: Conclusion and Future Work 34](#_Toc163503527)

[References 35](#_Toc163503528)

[List of abbreviations 39](#_Toc163503529)

# Chapter-1: Introduction

## 1.1. Background

As an important research direction in the fields of artificial intelligence and computer science, intention detection aims to enable computer systems to understand 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 [1]. 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 [2]. The application of these technologies promotes the progress of intention detection technology in various fields and provides a more accurate and intelligent user interaction experience for intelligent systems [3]. However, traditional intention detection methods are usually limited to single-modal data analysis and do not make full use of the rich information of multimedia data. In practical applications, users often use multiple forms of communication such as images and texts at the same time, and single-modal methods are difficult to fully understand the user's real intention.

In the era of digital multimedia, the field of intention detection is facing more complex and diverse user expressions. As a cutting-edge research method, multimodal data fusion, especially the fusion of image and text, provides a new idea to solve this problem, as shown in **Figure 1.1**. Baltrušaitis, et al in [4] use a survey-based approach to categorize and analyze various fusion strategies, learning paradigms, and evaluation methods in multimodal machine learning. They conclude that multimodality usually mainly contains three parts: representation, alignment, and fusion, among which representation and fusion are more important. Huang, et al in [5] propose an effective representation and fusion method based on attention mechanisms for end-to-end multimodal intent recognition. However, in the representation stage, the limitations of traditional methods in global context understanding are ignored. Inspired by the single-modal pre-training model, a series of multimodal pre-training works based on image-text data have emerged. Multimodal representation pre-training aims to use self-supervised learning paradigms, including contrastive learning, mask self-supervision, etc., to train on large-scale image-text correlation data. The model obtains universal and strong generalization ability of multimodal representation. Compared with a single modality, multimodal fusion contains more information. On the one hand, this information can complement each other. On the other hand, there is also the possibility of false data or contradictory phenomena. Therefore, how to correctly use and deal with the relationship and contradiction between each modality to improve the accuracy of intention detection has become one of the main research directions in this field [4].

图形用户界面, 网站

描述已自动生成

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

## 1.2. Motivation

Multimedia includes text, image, audio, video, and other forms of information expression. In recent years, the explosive growth of social media platforms has provided a wide range of dissemination platforms for multimedia content. Multimedia forms of information better meet the needs of users to express themselves, obtain information, and engage in interaction on social media. However, this expansion also brings challenges, especially in understanding and predicting user intentions in these digital spaces.

As one of the core technologies of human-computer interaction, intention detection aims to accurately identify the user's intention from their input to achieve more intelligent services. In social media, understanding whether a post is ironic, supportive, or critical through intention detection is crucial for improving recommendation systems, content moderation, and fostering healthier online communities. However, there are some potential weaknesses in traditional intention detection methods, such as only single modal data such as text or image are often considered, and the advantages of multimedia data are not fully utilized.

Existing research based on image-text information fusion mainly focuses on sentiment analysis and content classification, but there are still significant gaps in interpreting the subtle intentions behind user behavior. Based on this, this study proposes an intention detection method based on image and text fusion. Through this research, we can not only verify the feasibility and effectiveness of image and text information fusion technology in intention detection, but also make it more suitable for diverse user communication scenarios in the real world.

## 1.3. Problem Statement

The traditional intention detection method is mainly based on manual rules and infers the user's intention through a pre-defined rule set. Although this method is simple, it requires manual arrangement of the rule set, has limited coverage and is not flexible enough. With the development of deep learning technology, researchers have proposed many intention detection methods based on deep learning [6, 7]. These methods can automatically learn relevant patterns from massive data and are more efficient and accurate than manual rules. However, using only single modality for intention detection often cannot fully utilize the diversity of information, and it is difficult to effectively complete the task in the face of the diversity and complexity of input content. Therefore, how to effectively integrate and utilize multimodal data in a multimedia environment is one of the main research issues in this field.

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

Kruk, et al in [8] propose a model to capture the complex meaning multiplication relationship between image and text in multimodal Instagram posts. While this model integrated text and image information to detect intention, in the multimodal fusion stage, only the simple fusion strategy is used, that is, adds the two vectors, and the interaction information between text and images is ignored.

Maharana, et al in [9] introduce a late-fusion approach for integration of the video signal with the captions signal for intention Detection. Although it shows significant improvements with unimodal pre-trained models, the HERO used in the article is a pre-trained model only for video language, its performance on image and text data is not yet known.

Huang, et al in [5] develop a novel feature fusion method based on attention mechanism, which can recognize and utilize the importance of different modalities. The study designs complex strategies in feature fusion to reduce potential noise. However, it employs original pre-trained models, BERT for text and ResNet50 for images, to extract features. This approach restricts the model's comprehensive understanding of the global context.

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

## 1.4. Research Significance

In view of the abundance of image-text data on multimedia platforms, this study has important research significance for intent detection by integrating image and text information. First, the intention detection of image and text integration has important application value in social media analysis. By analyzing the intention of posts, we can understand the real thoughts of users, thereby providing personalized services such as advertising marketing. Secondly, many intelligent products are already well known. Integrating information from different media enhances intent detection accuracy, enabling intelligent systems to offer more flexible and natural interaction methods. This facilitates correct judgments or reasonable feedback on human intentions, thereby improving the quality and efficiency of human-computer interaction. Finally, in actual application scenarios, the proposed method can better adapt to data diversity, thereby improving the robustness and applicability of the intention detection system. Intention detection based on image-text fusion has become an important research direction in the field of artificial intelligence. This trend not only promotes the development of human-computer interaction technology, but also provides more powerful and intelligent application prospects for intelligent systems in various fields.

## 1.5. Research Questions

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

## 1.6. Research Objectives

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

## 1.7. Scope of the Study

The scope of the study revolves around enhancing intent detection in multimedia content through the fusion of image and text modalities. Traditional approaches relying solely on one modality often fail to capture the complexity of user intentions. The study aims to address this limitation by proposing a new method that integrates image and text data. The proposed method involves two key stages: multimodal representation and fusion. By leveraging multimodal pre-training technology and attention mechanism, which can capture more important information, to improve the accuracy of intent detection. The effectiveness of the proposed approach is validated through comparative experiments with a baseline model using a public multimodal intention dataset.

In this thesis, Chapter 1 presents the background and motivation for the research, emphasizing the importance and the challenges in detecting user intentions within multimedia content. Chapter 2 highlights the literature review, focusing on the advancements in intention detection and the evolution of multimodal pre-training models. We also discuss various methodologies and models that have been employed in the field in this part, providing a foundation for the proposed methodology. We detail the proposed methodology in chapter 3, describing a novel intention detection framework utilizing the CLIP model and cross-attention mechanism to fuse image and text data. In chapter 4, we present the dataset and evaluation metrics used, implementation details, experimental results, ablation studies, and analysis of the encoders' influence. Finally, we conclude the dissertation in chapter 5 by summarizing the findings, discussing the limitations of the current study, and suggesting directions for future research.

# Chapter-2: Literature Review

## 2.1. Intention Detection

Recently, with the rapid development and application of multimedia technology, users are now more inclined to express their intentions through multimodal data such as text and images. In fact, multimodal data contains richer information, and the accuracy of intention detection can be improved by learning from multimodal data. At present, intention detection based on image and text fusion has become a research hotspot in the field of artificial intelligence. In the early days, researchers used machine learning methods to detect intention. For example, Kuchlous, et al in [10] compared the effects of the bag-of-words model, TF-IDF and n-gram methods in short text intention analysis. Schuurmans, et al in [11] 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[12]. Yolchuyeva, et al in [13] presented a novel intention detection system which is based on a self-attention network and a Bi-LSTM. Obuchowski, et al in [14] proposed a novel approach to intention detection which involves combining transformer architecture with capsule networks. Chakraborty et al in [15] 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 in [16] 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 [17], multimodal technology can be used to analyze text and image information, identify users' emotional tendencies and expressions. In terms of humor detection [18], 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 in [8] proposed a model to capture the complex meaning multiplication relationship between image and text in multimodal Instagram posts. Maharana et al in [9] proposed a late-fusion approach for the integration of the video signal with the captions signal for intention detection. Other related work is summarized and highlighted in Table 2.1.

Table 2.1: Related work on intention detection

|  |  |  |  |
| --- | --- | --- | --- |
| **Title** | **Model/Method** | **Pros** | **Cons** |
| **Dictionary-based and Rule-based** | | | |
| Finding suggestions and buy wishes from product reviews [19] | Rules and Graphs | easy to realize.  fast operation speed | high workload poor stability |
| **Machine Learning** | | | |
| intention classification for dialogue utterances [20] | Naive Bayes bag-of-words | considering the order of words | cannot be used in new data |
| Short Text Intent Classification for Conversational Agent [21] | TF-IDF  N-gram | considering the importance of words | inability to accommodate large relation |
| Intent Detection through Text Mining and Analysis [22] | SVM  POS | Part-of-Speech considered | Average performance |
| **Deep Learning** | | | |
| User Intent Prediction in Information-seeking Conversations [23] | CNN | proving the effectiveness of CNN | CNN can only capture local semantic features |
| A RNN Contextual Approach to Intent Classification for Goal-oriented Systems [24] | RNN | capturing the features of the entire text | gradient explosion or disappearance for long text |
| Self-Attention Networks for Intent Detection [25] | Self-attention BILSTM | capturing long-range and multi-scale dependencies | low efficiency |
| Encoding syntactic knowledge in transformer encoder for intent detection and slot filling [26] | Transformer encoder-based | encode syntactic knowledge into  the model | only can be used for text |
| Intent recognition model based on sequential information and sentence features [27] | CNN  BILSTM  BERT | leverages contextual and semantic information within the text | require higher computational and storage resources |
| **Multimodal** | | | |
| An effective multimodal representation and fusion method for multimodal intent recognition [5] | Attention  BERT  Faster R-CNN | complementarity and consistency are considered | model structure is complex.  require higher computational and storage resources |

## 2.2. Multimodal Pre-training

With the gradual maturity of pre-training model technology in the field of natural language, multimodal pre-training models have gradually attracted attention, and a series of visual-language pre-training work has emerged. Vision-and-Language Pre-training VLP (Vision-and-Language Pre-training) [28] 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 [29] and ALIGN [30], this type of method uses contrastive learning for pre-training. They use cosine similarity to measure the distance between modalities and have demonstrated excellent performance in different fields. Li et al in [31] proposed a novel visual language pre-training framework ALBEF, which adds an intermediate amount of image-text contrast loss to enable the multimodal encoder to perform better cross-modal alignment. Yang et al in [32] 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. Li et al in [33] proposed BLIP, hoping to train a unified multi-modal pre-trained model to solve multi-modal understanding and generation tasks simultaneously. BLIP is a hybrid multi-modal encoder-decoder that can encode images or text in a single mode, image-based text coding and image-based text decoding. Recently, Meta AI He Kaiming's team launched the FILIP [34] multimodal pre-training model, which integrates the image-text double masking technology in MAE [35] 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 [36] and LCMERT [37] 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. Wang et al in [38] proposed SimVLM. Different from the general multi-modal pre-trained model using MLM, SimVLM uses the prefixLM method to preserve the visual language representation. Qi et al in [39] proposed Image BERT, 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. ViLT [40] 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, but there is still a certain gap between it and the CLIP twin-tower framework.

After analyzing representative methods in the field of intention detection in recent years, it can be concluded that the research content mainly includes three parts: feature extraction, multimodal representation, and multimodal fusion. However, most researchers only focus on the multimodal fusion part and propose some new methods, but the multimodal feature extraction and representation part is less considered and only the traditional single-modal feature extraction method is used. However, the development of multimodal large-scale pre-trained models gives us new ideas for feature extraction and representation.

# Chapter-3: Proposed Methodology

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

图示

描述已自动生成

Figure 3.1. Intention detection architecture diagram

## 3.1. Multimodal Representation

The quality of input features has an important impact on the prediction results of multimodal intention detection models. As early as the machine learning period, feature engineering determined the upper limit of learning. Better features mean you don’t need complex models to get excellent results. With the development of deep learning neural networks, the method of feature representation has also changed greatly. Currently, in multimodal intention detection, BERT and ResNet pre-trained models are mainly used to extract text and image features. However, BERT and ResNet are usually trained independently, this result in each model only understanding the information of its specific modality. This method may limit 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 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. As shown in the **Figure 3.2** 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. Image Encoder is used to extract image features and can adopt the latest proposed ViT-B/16 Transformer architecture.

日程表

描述已自动生成

Figure 3.2. Feature representation based on CLIP.

### 3.1.1. Image Encoder

Vision Transformer ViT-B/16 is used for image coding. It is an image classification model based on Transformer, where B represents the basic version, and 16 represents that the image is divided into 16×16 image blocks. Compared with traditional 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. The detail architecture is as shown in **Figure 3.3**. 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.

图示

描述已自动生成

Figure 3.3. Vision Transformer architecture diagram

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

Then, map input vectors to embedding vectors ​, add positional information to consider the sequence of inputs, W and b represent embedding parameters and bias terms, respectively:

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

Where, are the query, key, and value obtained by linear transformation, represent the embedding vectors, are different weight matrices,  is the dimensionality of the attention heads.

### 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. The detail architecture is as shown in **Figure 3.4**.

日程表

低可信度描述已自动生成

Figure 3.4. Text Transformer architecture diagram.

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

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:

## 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 [41], decision-level fusion [42] and hybrid fusion [43]. 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. **Figure 3.5** present the three different fusion strategies.

手机屏幕的截图

描述已自动生成

Figure 3.5. Three different fusion strategies

To extract deep features of different modalities and better integrate information between different modalities, this study adopts a feature-level fusion strategy to fuse image and text features based on a cross-modal attention mechanism. Different from the simple vector splicing method, based on Multimodal fusion with cross-modal attention mechanism refers to using the attention mechanism to adjust the attention between modalities 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. **Figure 3.6** show the process of Cross-attention calculation.

图示

描述已自动生成

Figure 3.6. Cross-attention calculation process.

We have an image representation and a text representation , where each and are feature vectors.

For the image and text , calculate Queries , Keys , and Values :

Text to Image Attention: calculate attention scores by taking the dot product of Image Key () and Text Query (), then apply SoftMax for normalization:

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

Here,  are weight matrices, is the dimensionality of the attention heads.

## 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.

Where W and b represent linear layer parameters and bias terms, respectively. We use 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.

is the total number of intentions, is the one-hot representation of the sample label, and represents the probability that the sample belongs to the i-th category.

# Chapter-4: Experimental Section

In this part, we first train and test the multimodal intent detection method based on image and text information fusion proposed in this research. Then 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) [44], 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" contain 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 metrics to measure the accuracy of the model, and its calculation method is shown in Formula 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:

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:

However, for datasets with an unbalanced number of positive and negative samples, the above evaluation metrics have certain flaws. In certain unbalanced samples, 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:

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 GPUs 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, we use CLIP to extract text and image features simultaneously, in which the image encoder and text encoder use ViT-B/16 and the transformer structure 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, 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 [45]. 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 in [46] propose 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.

BERT [47] 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.

ResNet-50 [48] 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. 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 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. This may be 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. Thanks to the development of large-scale pre-trained language models, text can usually 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. This indicates 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 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 4.1**, 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

Figure 4.1. 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 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.

# References

[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[2] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

[3] Chen, T., Liu, B., Yang, M., Liu, C., & Chang, M. C. (2018). The state of the art of deep learning in language processing. ACM Transactions on Information Systems (TOIS), 37(4), 1-43.

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

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

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

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

[8] 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.

[9] 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).

[10] 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.

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

[12] 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.

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

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

[15] 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.

[16] Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020). Efficient intention detection with dual sentence encoders. arXiv preprint arXiv:2003.04807.

[17] Dashtipour, K., Gogate, M., Cambria, E., & Hussain, A. (2021). A novel context-aware multimodal framework for persian sentiment analysis. Neurocomputing, 457, 377-388.

[18] 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).

[19] Ramanand, J., Bhavsar, K., & Pedanekar, N. (2010, June). Wishful thinking-finding suggestions and’buy’wishes from product reviews. In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text (pp. 54-61).

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

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

[22] Akulick, S., & Mahmoud, E. S. (2017). Intent detection through text mining and analysis.

[23] Qu, C., Yang, L., Croft, W. B., Zhang, Y., Trippas, J. R., & Qiu, M. (2019, March). User intent prediction in information-seeking conversations. In Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (pp. 25-33).

[24] Mensio, M., Rizzo, G., & Morisio, M. (2018, April). Multi-turn qa: A rnn contextual approach to intent classification for goal-oriented systems. In Companion Proceedings of the The Web Conference 2018 (pp. 1075-1080).

[25] Yolchuyeva, S., Németh, G., & Gyires-Tóth, B. (2020). Self-attention networks for intent detection. arXiv preprint ar\*\*v:2006.15585.

[26] 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).

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

[28] 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).

[29] 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.

[30] 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.

[31] 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.

[32] 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).

[33] 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.

[34] 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).

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

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., & Poria, S. (2018). Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-based systems, 161, 124-133.

[42] Liao, C. Y., Chen, R. C., & Tai, S. K. (2018, April). Emotion stress detection using EEG signal and deep learning technologies. In 2018 IEEE International Conference on Applied System Invention (ICASI) (pp. 90-93). IEEE.

[43] Nemati, S., Rohani, R., Basiri, M. E., Abdar, M., Yen, N. Y., & Makarenkov, V. (2019). A hybrid latent space data fusion method for multimodal emotion recognition. IEEE Access, 7, 172948-172964.

[44] 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).

[45] Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019, July). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the conference. Association for Computational Linguistics. Meeting (Vol. 2019, p. 6558). NIH Public Access.

[46] Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., Mao, C., Morency, L. P., & Hoque, E. (2020, July). Integrating multimodal information in large pretrained transformers. In Proceedings of the conference. Association for Computational Linguistics. Meeting (Vol. 2020, p. 2359). NIH Public Access.

[47] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, p. 2).

[48] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

# List of abbreviations

|  |  |
| --- | --- |
| **Abbreviations** | **Definition** |
| CLIP | Contrastive Language-Image Pre-training |
| HERO | Hierarchical Encoder for Video Language Omni-representation |
| BERT | Bidirectional Encoder Representations from Transformers |
| ResNet | Residual Network |
| TF-IDF | term frequency–inverse document frequency |
| SVM | supported vector machine |
| Bi-LSTM | Bi-directional Long Short-Term Memory |
| USE | Universal Sentence Encoder |
| POS | Part-of-Speech tagging |
| RNN | Recurrent neural network |
| CNN | convolution neural network |
| VLP | Vision-Language Pre-traning |
| ALIGN | A Large-scale ImaGe and Noisy-text embedding |
| ALBEF | ALign the image and text representations BEfore Fusing |
| TCL | Triple Contrastive Learning |
| BLIP | Bootstrapping Language-Image Pre-training |
| FILIP | Fine-grained Interactive Language-Image Pretrain |
| MAE | Masked Autoencoders |
| ViLBERT | Vision-and-Language BERT |
| LCMERT | Learning Cross-Modality Encoder Representations from Transformers |
| MLM | Masked Language Model |
| SimVLM | Simple Visual Language Model |
| ViLT | Vision-and-Language Transformer |
| NLP | Natural Language Processing |
| ViT | Vision Transformer |
| MIntRec | Multimodal Intent Recognition |
| MLP | multi-layer perceptron |
| MulT | Multimodal Transformer |
| MAG | Multimodal Adaptation Gate architecture |
| CV | Computer Vision |