IMAGE AND TEXT FUSION FOR INTENT DETECTION IN MULTIMEDIA

WANG MENG

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE (APPLIED COMPUTING)

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2024

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: WangMeng (I.C/Passport No: EE0343564)

Matric No: S2164723

Name of Degree: MASTER OF COMPUTER SCIENCE (APPLIED COMPUTING)

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Image and Text Fusion for Intent Detection in Multimedia

Field of Study:

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

IMAGE AND TEXT FUSION FOR INTENT DETECTION IN MULTIMEDIA

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-training 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 character 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.

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

GABUNGAN IMEJ DAN TEKS UNTUK PENGESANAN NIAT DALAM MULTIMEDIA

ABSTRAK

Dalam bidang interaksi manusia-komputer, pengesanan niat manusia merupakan masalah yang mencabar dan pautan utama untuk mencapai komunikasi tanpa halangan antara manusia dan mesin. Dengan evolusi pesat teknologi multimedia dan penggunaan meluas platform media sosial, pengesanan niat pengguna menjadi semakin mencabar. Pendekatan unimodal tradisional, terutamanya yang bergantung semata-mata pada maklumat teks atau visual, sering gagal menangkap selok-belok niat pengguna dalam kandungan multimedia. Untuk menangani had ini, gabungan modaliti imej dan teks menggunakan teknologi multimodal telah muncul sebagai penyelesaian yang menjanjikan untuk pengesanan niat. Berbanding dengan data mod tunggal seperti imej dan teks, data Multimodal boleh menawarkan maklumat yang lebih kaya untuk pengesanan niat pengguna yang lebih tepat. Pada masa ini, terdapat penyelidikan terhad mengenai pengesanan niat melalui gabungan imej dan teks. Kajian sedia ada tertumpu terutamanya pada penyepaduan ciri modal semasa gabungan ciri, menggunakan hanya model pra-latihan mod tunggal untuk pengekstrakan ciri. Walau bagaimanapun, kaedah ini merumitkan sistem, meningkatkan kos pengiraan, dan mungkin mengehadkan pemahaman kontekstual model, sekali gus menghalang penangkapan maklumat korelasi antara mod yang berkesan. Penyelidikan ini berusaha untuk mereka bentuk rangka kerja pengesanan niat yang baru untuk meningkatkan ketepatan dan keteguhan pengesanan niat. Kaedah yang dicadangkan merangkumi dua peringkat yang sama penting dalam perwakilan dan gabungan pelbagai mod. Dalam bahagian perwakilan, kami memperkenalkan model pra-latihan berskala besar multimodal untuk mengekstrak ciri teks dan imej secara serentak. Dalam bahagian gabungan, kami menggunakan kaedah gabungan berbilang peringkat Cross-modal berasaskan perhatian, yang boleh menggabungkan ciri imej dan teks daripada tahap token dan peringkat global. Untuk mengesahkan keberkesanan kaedah yang dicadangkan, kajian ini membangunkan rangka kerja pengesanan niat dan menjalankan pengesahan eksperimen pada dataset pengesanan niat. Keputusan eksperimen menunjukkan bahawa model Acc sebanyak 0.45-1.1% dan nilai F1 sebanyak 0.25-0.73% berbanding model multimodal asas lain, yang mengesahkan keberkesanan model dalam penyelidikan ini. Kami juga menggunakan eksperimen ablasi untuk mengesahkan kesahihan setiap modul model.

Kata Kunci: pengesanan niat, teknologi multimodal, Contrastive Language-Image Pra-Training (CLIP), representasi ciri, gabungan multimodal

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Suzan Jabbar Obaiys, whose guidance and support were invaluable throughout the entire process of this research. Your insight and encouragement have been crucial to my academic and personal growth.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

CLIP	:	Contrastive Language-Image Pre-training
SVM	:	Supported Vector Machine
USE	:	Universal Sentence Encoder
POS	:	Part-of-Speech tagging
RNN	:	Recurrent neural network
CNN	:	Convolution neural network
VLP	:	Vision-Language Pre-traning
MLM	:	Masked Language Model
MAE	:	Masked Autoencoders
ViLT	:	Vision-and-Language Transformer
NLP	:	Natural Language Processing
ViT	:	Vision Transformer
CV	:	Computer Vision
BERT	:	Bidirectional Encoder Representations from Transformers
TFIDF	:	term frequency–inverse document frequency

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 (LeCun & Bengio, 2015). 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 (Schmidhuber, 2015). 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 (Chen & Liu, 2018). However, traditional intention detection methods are usually limited to single-modality 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-modality 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 and Ahuja (2018) use a survey-based approach to categorize and analyze various fusion strategies, learning paradigms, and evaluation methods in multimodal machine learning. They conclude that multimodal usually mainly contains three parts: representation, alignment, and fusion, among which representation and fusion are more important. Huang and Ma (2023) 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-modality 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 (Baltrušaitis & Ahuja, 2018).

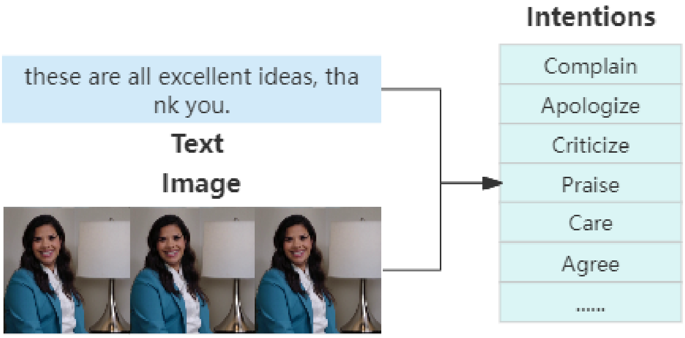

Figure 1.1: an example of intent detection based on text and image fusion

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 modality 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 (Wang & Wei, 2021). 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 and Lubin (2019) proposed 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 and Tran (2022) introduced 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-training models, the HERO used in the article is a pre-training model only for video language, its performance on image and text data is not yet known.

Huang and Ma (2023) developed 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-training models, BERT (Bidirectional Encoder Representations from Transformers) 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-modality 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-training 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

Multimodal intention detection has been widely used in many fields, and its practical value is huge. 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.

The multimodal intention detection algorithm based on image and text fusion has more advantages in robustness and accuracy, and has gradually become the mainstream of multimodal intention analysis research. From a technical point of view, such algorithms can capture the user's intention more comprehensively by combining image and text data. For example, in e-commerce platforms, product images and description texts uploaded by users together constitute a complete expression of intent. By fusing the data of these two modalities, the algorithm can more accurately understand the needs and preferences of users, so as to provide more accurate recommendation and search results. In addition, fusing multi-modal data can also enhance the robustness of the model, making it perform better in dealing with noise and incomplete data. This multi-modal fusion method greatly improves the accuracy and reliability of intention recognition, and has become a hot and mainstream of current research.

Cross-modality multi-layer attention mechanism can better capture the interaction information between modalities, which is of great significance for the study of multimodal machine learning. From a technical point of view, the cross-modal multi-layer attention mechanism can effectively identify and capture the subtle interaction information between modalities by establishing a hierarchical attention model between different modalities. For example, in the process of image and text fusion, cross-modal multi-layer attention mechanism can focus on key regions in the image and important words in the text, so as to extract more representative and relevant features. This not only improves the interpretability and controllability of the model, but also enhances the flexibility and accuracy of the model when dealing with complex scenes. Cross-modal multi-layer attention mechanism can capture the relationship between information in more detail by focusing on the interaction between modalities at different levels. For example, in video analysis, it is not only necessary to focus on visual information in image frames, but also to combine speech and text information in audio and captions. Through the multi-level attention mechanism, the model can understand the relationship between these modalities more comprehensively, so as to improve the accuracy and depth of the overall analysis. From an academic perspective, cross-modal multi-layer attention mechanism provides new theories and methods for multimodal machine learning research. Traditional multi-modal learning methods usually deal with the simple fusion of modalities, while the cross-modal multi-layer attention mechanism emphasizes the deep interaction and hierarchical information integration between modalities. The introduction of this method provides a new idea for exploring more complex and efficient multimodal fusion strategies in intent detection.

1.5 Research Questions

1. How can the multimodal (Vision-Language) pre-training model be leveraged for feature extraction and representation in the field of intent detection?
2. How to fuse image and text features from character-level and global-level based on attention mechanism?
3. How can the proposed method affect the accuracy of intent 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 fuse image and text features from character level and global level based on cross-modality multi-level attention mechanism.
3. To evaluate the performance of the proposed intent detection method 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

Intent detection is essentially a pattern recognition problem and has always been an important research content in the field of artificial intelligence. From the perspective of model structure, the related works can be roughly divided into traditional machine learning and deep learning. From the perspective of model input, we can divide the related works into single-modality and multimodal. A series of representative multimodal large-scale pre-training models have appeared since 2019 and achieved state of the art performance in many practical applications. According to the different input, it can be roughly divided into image-text, videotext and video-audio, in which the image-text is the most, because this type of content is the most common in the real world. How to effectively fuse multiple modalities has always been a key issue in the field of multimodal research. Multimodal fusion aims to integrate the information of different single modalities into a multimodal representation. Usually, the feature representation of single modalities is closely related to the fusion of multiple modalities, so most of the previous classification of multimodal fusion strategies is based on the stage of the fusion process in the overall flow.

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 and Kadaba (2020) compared the effects of the bag-of-words model, TF-IDF and n-gram methods in short text intention analysis. Schuurmans and Frasincar (2019) employed 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 (Louvan & Magnini, 2020). Yolchuyeva and Németh (2020) presented a novel intention detection system which is based on a self-attention network and a Bi-LSTM. Obuchowski and Lew (2020) proposed a novel approach to intention detection which involves combining transformer architecture with capsule networks. Chakraborty and Ohm (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 and Temčinas (2020) introduced 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 & Gogate, 2021), multimodal technology can be used to analyze text and image information, identify users' emotional tendencies and expressions. In terms of humor detection (Hasan & Lee, 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 and Lubin (2019) proposed a model to capture the complex meaning multiplication relationship between image and text in multimodal Instagram posts. Maharana and Tran (2022) 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 works on intention detection

Title	Model/Method	Pros	Cons
Dictionary-based and Rule-based			
Finding suggestions and buy wishes from product reviews (Ramanand & Bhavsar, 2010)	Rules and Graphs	easy to realize. fast operation speed	high workload poor stability
Machine Learning			
intention classification for dialogue utterances (Schuurmans & Frasincar, 2019)	Naive Bayes bag-of-words	considering the order of words	cannot be used in new data
Short Text Intent Classification for Conversational Agent (Kuchlous & Kadaba, 2020)	TF-IDF N-gram	considering the importance of words	inability to accommodate large relation
Intent Detection through Text Mining and Analysis (Akulick & Mahmoud, 2017)	SVM POS	Part-of-Speech considered	Average performance
Deep Learning			
User Intent Prediction in Information-seeking Conversations (Qu & Yang, 2019)	CNN	proving the effectiveness of CNN	CNN can only capture local semantic features
A RNN Contextual Approach to Intent Classification for Goal-oriented Systems (Mensio & Rizzo, 2018)	RNN	capturing the features of the entire text	gradient explosion or disappearance for long text
Self-Attention Networks for Intent Detection (Yolchuyeva & Németh, 2020)	Self-attention BILSTM	capturing long-range and multi-scale dependencies	low efficiency
Encoding syntactic knowledge in transformer encoder for intent detection and slot filling (Wang & Wei, 2021)	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 (Wu & Wang, 2024)	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 (Huang & Ma, 2023)	Attention BERT Faster R-CNN	complementarity and consistency are considered	model structure is complex. require higher computational and storage resources

2.2 Multimodal Representation

Modality refers to a certain type of information. Each source or form of information can be called a modality, mainly including text, image, audio and video. multimodal refers to the combination of two or more modalities, and multimodal learning establishes a model that can process multimodal information. In 2018, Baltrušaitis and Ahuja (2018) summarized the research status of multimodal learning and proposed five challenges for the future development of multimodal learning: representation, fusion, inter-modal mapping, modal alignment, and collaborative learning.

The multimodal representation is the basis of multimodal learning, and a good representation can greatly improve the performance of the model. There are many challenges in representing multimodal data, including how to obtain more effective representations, how to combine data from heterogeneous sources, and how to deal with different levels of noise. Bengio and Courville (2013) proposed the following properties of a good representation: smoothness, temporal and spatial coherence, sparsity, and natural clustering.

From the perspective of multimodal feature representation, existing approaches can generally be divided into two paradigms: Coordinated representations and Joint representations. Coordinated representations also known as aligned or correspondence-based representations, involve learning separate representations for each modality in such a way that these representations can be easily aligned or compared. The key idea is to project data from different modalities into a shared embedding space where similar concepts from different modalities are close to each other. Coordinated representations typically use separate encoders for each modality. For example, a convolutional neural network (CNN) might be used to encode images, while a Transformer-based model could encode text. Coordinated representations typically use separate encoders for each modality. Joint representations also known as integrated representations, involve learning a single, unified representation that directly combines information from multiple modalities. This approach aims to capture the interactions and dependencies between modalities more holistically within a single representation. Unlike coordinated representations, joint representations typically use a unified model that simultaneously processes multiple modalities. This might involve concatenating features from different modalities early in the model and processing them together. Joint representations are particularly effective at capturing complex interactions between modalities. Therefore, in this research, we use multimodal pre-training model to achieve multimodal representation through transfer learning method.

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. VLP (Vision-and-Language Pre-training) (Chen & Ding, 2020) refers to a universal representation of cross-modality 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 & Kim, 2021) and ALIGN (Jia & Yang, 2021), 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 and Selvaraju (2021) 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-modality alignment. Yang and Duan (2022) proposed TCL with triple contrastive learning by leveraging cross-modality and intra-modality self-supervision. TCL further considers intra-modality supervision to ensure that the learned representation is also meaningful in each modality, which facilitates cross-modality alignment and joint multimodality embedding learning. Li and Li (2022) proposed BLIP, hoping to train a unified multimodality pre-training model to solve multimodality understanding and generation tasks simultaneously. BLIP is a hybrid multimodality 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 (Li & Fan, 2023) multimodal pre-training model, which integrates the image-text double masking technology in MAE (Baade & Peng, 2022) and effectively improves the efficiency of model pre-training compared with CLIP.

The fusion coding framework uses the Transformer mechanism for cross-modality fusion. ViLBERT (Lu & Batra, 2019) and LCMERT (Tan & Bansal, 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. Wang and Yu (2021) proposed SimVLM. Different from the general multimodality pre-training model using MLM, SimVLM uses the prefixLM method to preserve the visual language representation. Qi and Su (2020) 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 (Kim & Son, 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, but there is still a certain gap between it and the CLIP twin-tower framework.

Some of these multimodal pre-training models are trained based on massive datasets by professional Artificial Intelligence Organization and have been used in many practical tasks through transfer learning. For example, in the Sentiment Classification task, the author extends the popular Vision-and-Language Transformer (ViLT) to process more complex text inputs than image captions, improving performance of model while having minimal impact on training and inference efficiency (Chochlakis & Srinivasan, 2022). In the Retrieval task, the author explores the zero-shot image classification and retrieval ability of CLIP in artworks domain (Baldrati & Bertini, 2022). As we can see from the recent studies in video summarization task (Zhao & Gong, 2022) and intent discovery task (Maharana & Trank, 2022), multimodal pre-training models have shown great performance in many tasks compared with traditional methods.

2.3 Multimodal Fusion

Although the deep learning technology based on single modality such as text, speech and image has made remarkable progress, due to the contradiction between the diversity of intention expression in real life and the limitations of single-modal data, the intention detection based on multimodal data can better meet the needs of actual scenes. Compared with single-modal intent detection, multimodal intent detection also needs to consider how to effectively fuse the features of different modalities, which is also one of the unique technical problems in the field of multimodal technology. In recent years, relevant researchers have been committed to effectively using the inter-modal consistency features and intra-modal diversity features in multimodal data to establish reasonable classification models, so as to obtain classification results with both robustness and accuracy. In general, the existing multimodal fusion strategies can be divided into three types, which are feature-level fusion, decision-level fusion and hybrid fusion.

Feature-level Fusion, also known as Early Fusion, usually restructures features extracted from different modalities into new feature representations by concatenation and weighting operations. In the early feature layer fusion, the features extracted from different modalities are concatenated in the same dimension, and then the feature dimension is reduced, and finally input into the classifier for category analysis. For example, Pérez-Rosas and Mihalcea (2013) extracted the features of speech, text and image data respectively, and then concatenate the features, and finally use the support vector machine for classification. With the development of deep learning, the feature extraction of single modality has changed from traditional manual feature extraction to automatic extraction by using deep neural network for training, and the way of feature fusion has also developed from simple splicing to the use of various networks for deep fusion. For example, Zadeh and Liang (2018) designed a memory fusion network, which can not only fuse feature representations between different modalities, but also model in the time dimension to obtain feature representations of context information. Tsai and Bai (2019) proposed a Transformer-based multimodal feature extraction network, which uses the attention mechanism to learn the feature representation between modalities, so as to enhance the representation ability of the fused features to the target information.

Decision-level Fusion, also known as Late Fusion, usually builds different models for multiple modalities to deal with the features extracted by a single modality, and then uses the fusion technique to fuse the model output results of different modalities to obtain the final category analysis results. For example, Wang and Guan (2012) used different feature extraction methods to extract features from data of different modalities, then used hidden Markov model to model the prediction of single modality, and finally used Kernel canonical correlation analysis to fuse the prediction results. Compared with feature-level fusion, decision-level fusion can adopt targeted feature extraction methods and classifiers for the data characteristics of different modalities, and the fusion results can better reflect the differences between modalities. However, the extraction of consistent features that are common between multimodal data is often insufficient.

Most of the recent research focuses on hybrid fusion methods. Different from early fusion and late fusion, hybrid fusion no longer emphasizes the stage of the fusion process, but focuses on the fusion of multimodal data features at different levels, which not only obtains the consistency between multimodal data, but also covers the uniqueness. For example, the deep neural network and training method based on Attention mechanism proposed by Harish and Sadat (2020). perform multi-level fusion of speech, image and text modalities, which can simultaneously have the advantages of feature-level fusion and decision-level fusion.

Multimodal fusion can make full use of the complementary information in multimodal data, which has obvious advantages for multi-category classification tasks, and different multimodal fusion methods can be applied to different scenarios. At present, multimodal fusion methods are applied in different scenarios such as visual question answering, behavior recognition, image caption generation, emotion recognition, and emotion analysis, and the fusion methods are also very different in different application scenarios.

In visual question answering task, Lin and RoyChowdhury (2015) proposed a bilinear CNN model, which uses two CNNS to extract image features, and multiplies the outer product of the obtained image features to obtain a bilinear vector, and then performs image classification. This model structure can model local features, so it shows the advantage of fine-grained in image classification. Bilinear models have achieved excellent performance in visual tasks such as semantic segmentation, image fine-grained recognition and face recognition. However, because the bilinear features are high-dimensional, the feature dimension is too high to facilitate classification calculation. Therefore, Gao and Beijbom (2016) proposed the compact bilinear pooling method, which has the same discrimination ability as the bilinear model, but the feature dimension of this method is greatly reduced. Fukui and Park (2016) proposed a multimodal compact bilinear pooling model to fuse multimodal features and promote the interaction between vision and language in visual question answering task.

In the task of image caption generation, Cho and Courville (2015) proposed an encoder-decoder architecture of attention mechanism, which obtained more interaction information by fusing two different modal features, so as to improve the effect of image caption generation. In order to select the most relevant image regions for word prediction at each decoding stage, Xu and Ba (2015) introduced and added an attention mechanism model, which can learn to describe the content of the image, and extended the attention mechanism to obtain soft attention and hard attention mechanisms respectively, and applied these two attention mechanisms to the decoder of LSTM. However, the codec only needs a small amount of visual information in the image to predict the non-visual information. Therefore, Lu and Xiong (2017) adopted an adaptive attention model that dynamically focuses on image regions at each decoding stage and is able to automatically decide when to look at the image and generate the next word.

In the action recognition task, Wang and Gao (2017) proposed a temporal and spatial two-stream 3D convolutional fusion network, which can recognize human actions in videos. The network decomposes the video into spatial and temporal frames, and fuses the spatial-temporal two-stream pyramid pooling with LSTM by taking a set of frame sequences as input. Finally, the softmax function is used to classify the behavior. Since efficient spatial-temporal representation plays a crucial role in video understanding, Zheng and An (2018) proposed a new relation network based on temporal pyramid pooling, which can extract multi-scale high-level features of sampling frames, and then connect the features of segments of the same length to infer the relationship of features in the same length segment. Finally, a comprehensive prediction is made for different relationships. Because video has high dimensionality, contains rich behavior information, and has different motion states, it is difficult for traditional RNN to capture complex action information. Therefore, YAO and JIANG (2023) proposed a method based on graph convolution and self-attention graph pooling to solve this problem.

In the scenario of emotion recognition, Gera and Balasubramanian (2020) proposed a model based on spatial channel attention network and context information complementary, which can eliminate redundant features. Mittal and Guhan (2020) proposed a context-aware multimodal emotion recognition algorithm, which mainly combines three aspects of context relations. One is emotion recognition based on multiple modalities, the other is to collect semantic context from the input image and encode this information using CNN based on self-attention mechanism. The third is the use of depth map pooling for modeling, which improves the accuracy of emotion recognition. Since the introduction of attention mechanism cannot significantly improve the model training performance in the current multimodal emotion recognition task, Verma and Wang (2020) used LSTM and tensor-based convolutional network to dynamically model the intra-modal and inter-modal interactions, and proposed a deep network with high-order common sequence and unique sequence information. The network can encapsulate the temporal granularity between different modalities, so as to be more accurate in sentiment classification. Ben-Younes and Cadene (2017) proposed a Tucker decomposition method based on multimodal tensors, which effectively parameterizes bilinear fusion between visual and textual modalities.

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-modality feature extraction method is used. However, the development of multimodal large-scale pre-training 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-training 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.

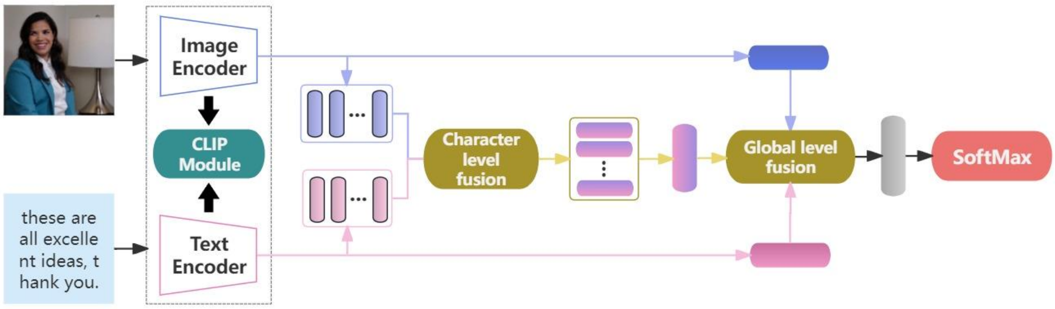

Figure 3.1: intention detection architecture diagram

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-training 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-modality 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-modality scheme and shows strong transfer ability. Moreover, a single multimodal pre-training 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-training 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-training 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.

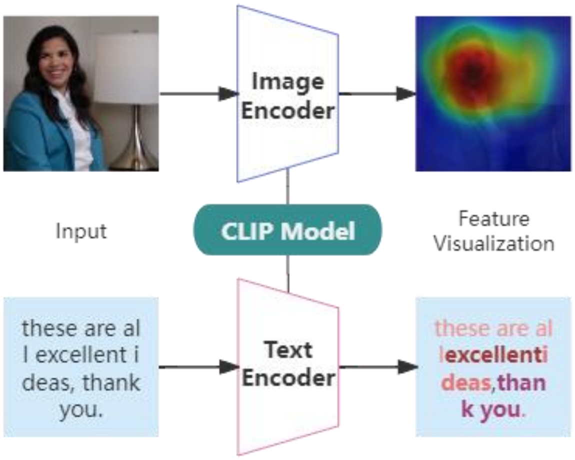
Image Encoder
Input
CLIP Model
Feature Visualization
these are al l excellent i deas, thank you.
Text Encoder
these are al lexcellenti deas,than k you.

Figure 3.2: feature representation based on CLIP

The CLIP model is a powerful tool for multimodal understanding and feature representation. One of its primary strengths is its robust zero-shot learning capability. Unlike traditional models that require task-specific fine-tuning, CLIP can perform a wide range of tasks directly out of the box. This is achieved by leveraging its training on a vast dataset of image-text pairs, enabling the model to understand and align visual and textual data in a shared embedding space. This allows CLIP to recognize and categorize new images based on textual descriptions, even if it has never seen those specific categories during training. This zero-shot ability significantly enhances the model's flexibility and applicability across different domains without the need for extensive labeled data. Another notable strength of CLIP is its ability to perform cross-modal retrieval tasks with high accuracy. By aligning image and text embeddings, CLIP enables efficient and effective retrieval of relevant images for a given text query. This cross-modal alignment is facilitated by the contrastive learning objective, which ensures that semantically similar images and texts are mapped close to each other in the embedding space. In addition, CLIP's use of powerful Transformer architectures for both image and text encoders contributes to its ability to capture rich and nuanced representations of multimodal data. This makes CLIP particularly useful for applications such as content-based image retrieval, visual question answering, and generating textual descriptions for images, thereby bridging the gap between visual and textual information and enhancing the overall user experience in multimodal interactions.

3.1.1 Image Encoder

The image encoder in the CLIP model is a critical component designed to convert visual data into a fixed-dimensional embedding that can be aligned with textual data embeddings. CLIP employs either a Vision Transformer (ViT) or a ResNet architecture for this purpose. The Vision Transformer (ViT) represents a transformative approach to computer vision by leveraging the power of transformers, which have traditionally been used in natural language processing (NLP). Introduced by the team at Google Research, ViT shifts from the convolutional neural networks (CNNs) paradigm to transformer-based architectures, bringing significant advancements in image recognition and classification tasks. Unlike CNNs, which rely on convolutional layers to extract local features and gradually build a hierarchical understanding of images, ViT processes an image as a sequence of patches, treating each patch similarly to tokens in NLP. This innovative methodology allows ViT to capture global context more effectively, which is particularly beneficial for complex image understanding tasks.

In ViT, an image is first divided into a grid of fixed-size patches, and each patch is flattened into a vector. These vectors are then linearly embedded and combined with positional encodings to retain spatial information. The resulting sequence of patch embeddings is fed into a standard transformer encoder, which consists of multi-head self-attention layers and feed-forward networks. The transformer architecture excels at capturing long-range dependencies and contextual information across the entire image, overcoming the locality constraint of CNNs. ViT has demonstrated remarkable performance on various benchmark datasets, often surpassing traditional CNNs, especially when pre-trained on large-scale datasets and fine-tuned on specific tasks. ViT's success heavily depends on the availability of extensive computational resources and large datasets, as its training is more data-hungry and computationally intensive compared to conventional CNNs. This also endow the CLIP model with the excellent transfer learning ability in practical application.

Vision Transformer ViT-B/16 is used for image coding in this method. 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.

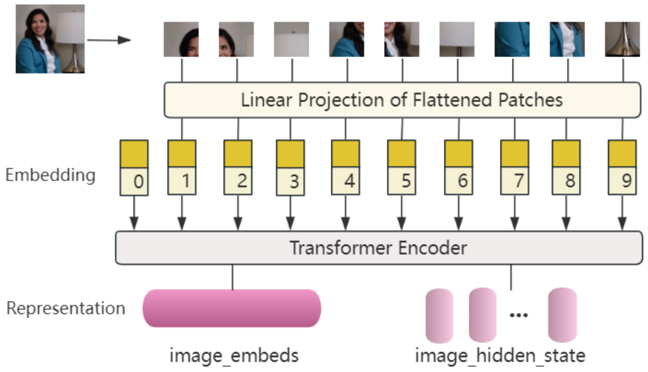

Figure 3.3: Vision Transformer architecture diagram

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

$$X = [x_1, x_2, \cdots, x_N] \tag{3-1}$$

Then, map input vectors x_i to embedding vectors z_i, add positional information to consider the sequence of inputs. Since Transformers are permutation invariant and lack the inherent spatial information that CNNs possess, positional encodings are added to the patch embeddings to retain the spatial structure of the image. These encodings help the model understand the relative positions of the patches within the image. W and b represent embedding parameters and bias terms, respectively:

$$z_i = W_{embed} \cdot x_i + b_{embed} + PositionalEmbedding(i) \tag{3-2}$$

In Transformer Encoder, calculate attention weights and output using self-attention mechanism. The sequence of patch embeddings, enriched with positional encodings, is fed into a standard Transformer encoder. This encoder consists of multiple layers of multi-head self-attention and feed-forward neural networks, each followed by layer normalization and residual connections:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3-3}$$

$$AttentionOutput = Attention\big(W_Q \cdot Z, W_K \cdot Z, W_V \cdot Z\ \big) \tag{3-4}$$

$$Q = W_Q \cdot Z \tag{3-5}$$

$$K = W_K \cdot Z \tag{3-6}$$

$$V = W_V \cdot Z \tag{3-7}$$

Where, Q, K, V are the query, key, and value obtained by linear transformation, z represent the embedding vectors, $W_{Q_I}, W_{K_I}, \cdots$ are different weight matrices, d_k is the dimensionality of the attention heads.

In general, after passing through the Transformer encoder, the sequence of embeddings is reduced to a single embedding by taking the output corresponding to a special classification token (CLS token) prepended to the sequence. This final embedding is then fed into a classification head, typically a multi-layer perceptron (MLP), to produce the final class probabilities. In this research, we use special classification token as global level feature output and use hidden state as character level feature output.

3.1.2 Text Encoder

The text encoder in the CLIP model plays a key role in mapping textual data into a shared embedding space with visual data. This encoder is based on the Transformer architecture, specifically leveraging a variant similar to the BERT model. The Transformer architecture is famous for its self-attention mechanisms, excels in capturing the intricate relationships and contextual dependencies within text. In the context of CLIP, the text encoder processes input text, such as captions or descriptions, and converts them into fixed-dimensional embeddings that can be aligned with the embeddings produced by the image 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. The Masked Self-Attention Transformer is a fundamental architecture in modern deep learning, particularly in the realm of natural language processing (NLP). This variant of the Transformer model, introduced in seminal works such as BERT, utilizes masked self-attention mechanisms to enhance contextual learning and language understanding. The core idea behind masked self-attention is to enable the model to predict missing or masked tokens within a sequence, thereby forcing it to learn deep contextual relationships among words. This is achieved by randomly masking a certain percentage of input tokens during training and then training the model to predict these masked tokens based on the surrounding context.

In a Masked Self-Attention Transformer, the input sequence is first tokenized and then passed through an embedding layer to generate dense vector representations for each token. Positional encodings are added to these embeddings to retain information about the sequence order. During the self-attention computation, the model applies masks to the token embeddings, which prevents certain tokens from attending to future tokens in the sequence. This masking ensures that the model only utilizes the available context, rather than "cheating" by looking ahead at the unmasked tokens. The masked self-attention layers enable the model to focus on relevant parts of the context while disregarding the masked positions, effectively learning bidirectional representations that capture the meaning and dependencies of the words within the text. These rich representations are then used for various downstream tasks such as language modeling, text classification, and more, demonstrating the powerful capabilities of Masked Self-Attention Transformers in understanding and generating human language.

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.

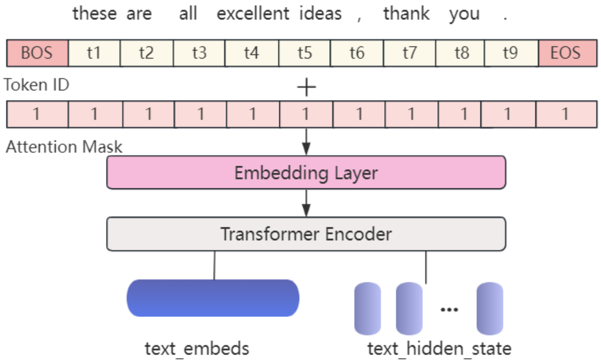

Figure 3.4: Text Transformer architecture diagram

The input text, such as a caption or a description, is first tokenized. Tokenization involves splitting the text into smaller units, typically words or subworlds, which can be efficiently processed by the model. These tokens are then converted into numerical representations through an embedding layer. Tokenize input text into tokens, embedding layer transforms each token t_i into an embedding vector e_i:

$$T = [t_1, t_2, \cdots, t_N] \quad (3-8)$$

$$E = [e_1, e_2, \cdots, e_N] \quad (3-9)$$

Transformers do not inherently understand the order of tokens in a sequence. To incorporate this crucial information, positional encodings are added to the token embeddings. These encodings allow the model to take into account the position of each token within the sequence, which is essential for understanding the syntax and semantics of the text.

Apply multiple transformer encoder blocks to process the token embeddings and positional encodings, the encoder structure is the same as in image encoder, each layer consists of a multi-head self-attention mechanism and a feed-forward neural network. The self-attention mechanism enables the model to weigh the importance of different tokens relative to each other, capturing nuanced relationships and contextual dependencies within the text. The feed-forward network further processes these relationships, enhancing the representation of the text. In the self-attention mechanism, the model computes attention scores for each token with respect to all other tokens in the sequence. This allows the model to focus on relevant parts of the context when processing each token. The self-attention outputs are then aggregated to form a comprehensive understanding of the text. After processing through the Transformer layers, the output is a sequence of contextually rich embeddings. A special token is used to aggregate the information of the entire text sequence and the hidden state is used to show the information of every token. This final embedding represents the entire text in a high-dimensional space.

$$E_{pos} = E + PositionalEncoding(1,2,\cdots,N) \qquad (3-10)$$

$$E_{encoded} = TransformerEncoder\left(E_{pos}\right) \qquad (3-11)$$

$$TextEmbedding = MeanPooling(E_{encoded}) \qquad (3-12)$$

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.

Multimodal fusion refers to the integration of information from different types of data modalities, such as text, image, audio, and video, to improve the performance of machine learning models on various tasks like classification, retrieval, and generation. Common strategies for multimodal fusion can be broadly categorized into feature-level fusion, decision-level fusion and hybrid fusion. Feature-level, also known as early fusion, combines raw data or feature representations from different modalities at an early stage. The fused features are then processed by subsequent layers of the model. This approach leverages the complementary nature of different modalities from the outset, enabling the model to learn joint representations that capture interactions between modalities. Decision-level, or late fusion, involves integrating the outputs of unimodal models at a later stage, often by combining their predictions. This can be done using simple techniques like averaging or voting, or more complex methods like stacking or ensemble learning. Hybrid fusion combines elements of both early and late fusion, integrating features at multiple stages. This allows the model to capture both low-level correlations and high-level interactions between modalities.

In general, 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.

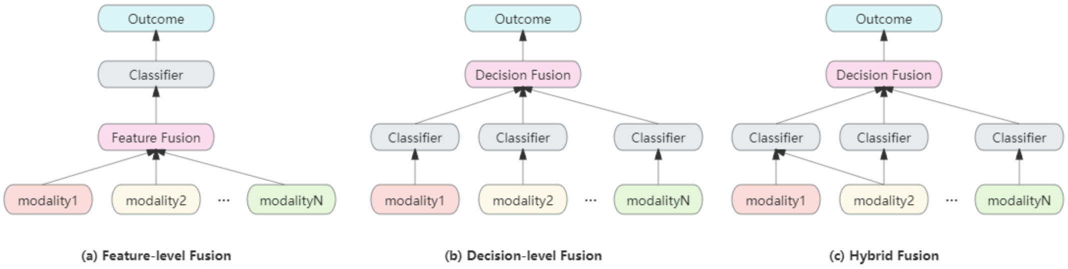

Figure 3.5: three different fusion strategies

The attention mechanism, originally developed for natural language processing, has become a powerful tool for multimodal fusion due to its ability to dynamically focus on relevant parts of the input data. Two key types of attention mechanisms used in multimodal fusion are self-attention and cross-attention. Self-attention, also known as intra-modal attention, computes attention weights within a single modality. This mechanism allows the model to dynamically adjust the importance of different parts of the input data, enhancing the representation of the modality. In the context of multimodal fusion, self-attention can be applied independently to each modality before combining their representations. For example, a model might apply self-attention to both text and image features separately to highlight the most informative parts of each modality. Cross-attention, or inter-modal attention, computes attention weights across different modalities. This allows the model to align and integrate information from multiple sources. For instance, cross-attention can be used to determine which parts of an image are most relevant to a given text description. In practice, cross-attention mechanisms can be implemented in various ways, such as attending from one modality to another or using a joint attention mechanism that simultaneously attends to both modalities.

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-modality attention mechanism. Different from the simple vector splicing method, based on Multimodal fusion with cross-modality 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-modality 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.

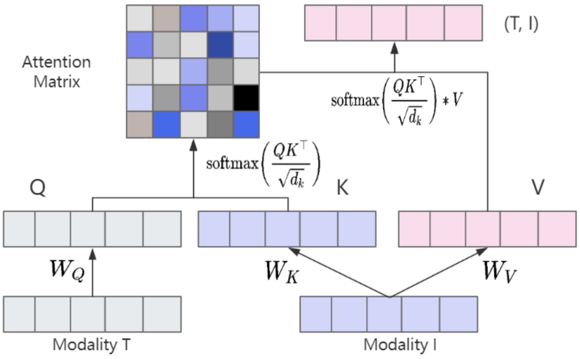
Attention Matrix
(T, I)
$\mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) * V$
$\mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)$
Q
K
V
W_Q
W_K
W_V
Modality T
Modality I

Figure 3.6: cross-attention calculation process

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

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

$$Q_I = \mathrm{I} \cdot \mathrm{W}_{Q_I},\quad \mathrm{K}_I = \mathrm{I} \cdot \mathrm{W}_{K_I},\quad \mathrm{V}_I = \mathrm{I} \cdot \mathrm{W}_{\mathrm{V}_I} \tag{3 - 13}$$

$$\mathrm{Q}_T = \mathrm{T} \cdot \mathrm{W}_{Q_T},\quad \mathrm{K}_T = \mathrm{T} \cdot \mathrm{W}_{K_T},\quad \mathrm{V}_T = \mathrm{T} \cdot \mathrm{W}_{V_T} \tag{3 - 14}$$

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

$$Attention_{T \to I} = softmax\left(\frac{\mathrm{Q}_T \cdot \mathrm{K}_I^T}{\sqrt{d_k}}\right) \tag{3 - 15}$$

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

$$Output_{T \to I} = Attention_{T \to I} \cdot \mathrm{V}_I \tag{3 - 16}$$

Here, $W_{Q_I}, W_{K_I}, \cdots$ are weight matrices, d_k is the dimensionality of the attention heads.

In the feature-level multimodal fusion strategy, the application of cross-attention mechanisms, enhancing the feature fusion ability, enable models to effectively integrate and process diverse data sources. By leveraging self-attention in representation stage and cross-attention in fusion stage, these models can dynamically focus on relevant parts of the input data, capturing complex interactions and dependencies between modalities. This results in more accurate, robust, and contextually aware representations, driving advancements in intention detection based on image and text fusion.

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.

$$Softmax(Z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \qquad (3-17)$$

$$\hat{y} = \mathrm{softmax}(W * MLP(h) + b) \qquad (3-18)$$

Where W and b represent linear layer parameters and bias terms, respectively. We use Cross Entropy as the loss function, Cross entropy is a widely used loss function in machine learning, which is an important concept in Shannon information theory and is mainly used to measure the difference in information between two probability distributions. In the classification tasks of this research, cross entropy loss is used to evaluate how well the predicted probabilities match the actual class labels.

$$\mathrm{Loss} = -\sum_{i=1}^{n} y_i \cdot \log \hat{y}_i \qquad (3-19)$$

n is the total number of intentions, y_i is the one-hot representation of the sample label, and $\hat{y}_i$ 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) (Zhang & Xu, 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 modalities 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 4.1, we split training, validation, and testing sets in 6:2:2. The detailed statistics are shown in Table 4.2.

Table 4.1: 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 4.2: 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.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \qquad (4-1)$$

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.

Like 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:

$$Precision = \frac{TP}{TP + FP} \qquad (4-2)$$

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:

$$Recall = \frac{TP}{TP + FN} \qquad (4-3)$$

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:

$$F1 = \frac{2 * P * R}{P + R} \qquad (4-4)$$

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-modality 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-training CLIP weights are used as the initial weights of the image encoder and text encoder, and the weights in the cross-modality attention module and MLP classifier are randomly initialized. Other main hyperparameters are shown in the Table 4.3. The hyperparameter settings are mainly determined through observation results and based on prior knowledge.

Table 4.3: 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

In this research, we use gradient descent to minimize the loss function, ensuring that our model parameters are optimized to achieve the best possible performance on our dataset. Specifically, we employ the mini-batch gradient descent approach, which balances the computational efficiency of stochastic gradient descent with the stability of batch gradient descent. This method allows us to update our model parameters iteratively using small batches of data, thus accelerating the convergence while maintaining robustness against noisy gradients.

We also incorporate momentum to help accelerate the convergence of the gradient descent algorithm and to mitigate oscillations in the parameter updates. By adding a fraction of the previous update to the current update, momentum allows the algorithm to gain faster convergence rates, especially in scenarios involving high curvature or noise in the data.

Furthermore, to ensure that our model adapts effectively to varying data distributions and achieves faster convergence, we leverage adaptive learning rate techniques such as the Adam optimizer. Adam combines the advantages of both AdaGrad and RMSProp by maintaining per-parameter learning rates that are adapted based on the first and second moments of the gradients. This results in an algorithm that is both efficient and effective in handling sparse gradients and non-stationary objectives.

By integrating these advanced optimization techniques, we aim to train our model’s parameters effectively, reduce the overall training time, and enhance the model's ability to generalize to new data. This common model optimization approach to minimizing the loss function via gradient descent plays a crucial role in the success of our intent detection model, ensuring it performs accurately and robustly across the above four evaluation metrics.

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 modality learning models were selected for comparison with the method:

MulT (Tsai & Bai, 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 and Hasan (2020) 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 (Kenton & Toutanova, 2019) is a classic pre-training NLP model that adopts the Transformer architecture to learn universal language representations.

ResNet-50 (He & Zhang, 2016) is a pre-training 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-modality 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-modality 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 4.4: results for multimodal intent detection

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 4.4 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. We can see from the above table, the proposed method achieves the highest values in ACC, F1 and P three evaluation metrics compared with baseline models, in which ACC is improved by 0.22, P is improved by 0.08. This is mainly because we comprehensively consider the interactive information between image and corresponding text and the feature fusion of character level and global level. Secondly, from the perspective of input modalities, the results of multimodal models are generally better than the results of single-modality 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-training 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-modality 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 4.5, 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 splicing method to fuse features.

Table 4.5: 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 4.6:

Table 4.6: 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.

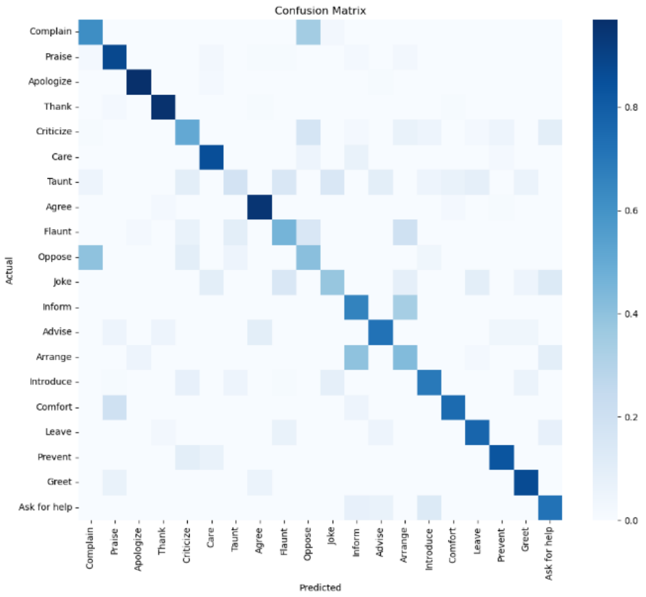

Figure 4.1: confusion matrix of test results

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: DISCUSSION AND CONCLUSION

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-training 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. There are other forms of modality in the dataset, such as audio and video, which may provide additional information. Incorporating these modalities can enrich the dataset, allowing the model to learn from multiple sources of information and improve its performance. For example, audio data can provide insights into tone and emotion, while video data can offer temporal and spatial dynamics that are not captured by static images. 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. The size and diversity of training samples are also crucial to enhancing model robustness and performance. In this research, the quantity of total samples used is relatively small compared to the number of labels. Increasing the number of training samples can lead to better model generalization and performance, particularly when dealing with a large variety of labels. Diverse training samples help the model to understand and learn from different contexts and variations, making it more adaptable to real-world scenarios. It is common for some modalities to be missing in real-world multimedia information, which can affect the robustness of the model. Handling missing modalities effectively is essential for the practical application of the model. We will explore more effective strategies such as data imputation, multi-task learning, or designing the model to be flexible to missing data in our future work to help mitigate this issue. Ensuring the model can work well even when some modalities are absent will enhance its usability and reliability in diverse environments. In the task of image-text multimodal intention detection, usually the dataset only provides the text and image uploaded by the user, and does not consider more context information. For the subsequent development, more information can be introduced from both internal and external perspectives. From the internal perspective, some user profile features can be added, such as the user's age, release location, release time, etc. These features can be interacted with the text content and image to give the model more information, and this information can further help the model to carry out more effective intention analysis and improve the effect of the model. On the other hand, from the external perspective, knowledge graph can be introduced. Since the content of text and image uploaded by users usually includes some information related to entities, it is often impossible to have a deep understanding of these information without introducing entity information. Some entity information in the text and image can be more fully used and understood, and the effect of the model can be further improved by introducing this entity information. In conclusion, addressing these limitations is of great significance for the practical application of the model, also is a possible research direction in the future.

REFERENCES

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

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

Baldrati, A., Bertini, M., Uricchio, T., & Del Bimbo, A. (2022, May). Exploiting CLIP-based multimodal approach for artwork classification and retrieval. In International Conference Florence Heri-Tech: The Future of Heritage Science and Technologies (pp. 140-149). Cham: Springer International Publishing.

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.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1798-1828.

Ben-Younes, H., Cadene, R., Cord, M., & Thome, N. (2017). Mutan: Multimodal tucker fusion for visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2612-2620).

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

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.

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).

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.

Cho, K., Courville, A., & Bengio, Y. (2015). Describing multimedia content using attention-based encoder-decoder networks. IEEE Transactions on Multimedia, 17(11), 1875-1886.

Chochlakis, G., Srinivasan, T., Thomason, J., & Narayanan, S. (2022). VAuLT: Augmenting the vision-and-language transformer for sentiment classification on social media. arXiv preprint arXiv:2208.09021.

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

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847.

Gao, Y., Beijbom, O., Zhang, N., & Darrell, T. (2016). Compact bilinear pooling. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 317-326).

Gera, D., & Balasubramanian, S. (2020). Affect expression behaviour analysis in the wild using spatio-channel attention and complementary context information. arXiv preprint arXiv:2009.14440.

Harish, A. B., & Sadat, F. (2020, April). Trimodal attention module for multimodal sentiment analysis (student abstract). In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 10, pp. 13803-13804).

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).

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).

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.

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.

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).

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.

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.

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.

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

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.

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.

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).

Lin, T. Y., RoyChowdhury, A., & Maji, S. (2015). Bilinear CNN models for fine-grained visual recognition. In Proceedings of the IEEE international conference on computer vision (pp. 1449-1457).

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.

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.

Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 375-383).

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).

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).

Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emoticon: Context-aware multimodal emotion recognition using frege's principle. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14234-14243).

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).

Pérez-Rosas, V., Mihalcea, R., & Morency, L. P. (2013, August). Utterance-level multimodal sentiment analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 973-982).

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.

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).

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.

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.

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).

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

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

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

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.

Verma, S., Wang, J., Ge, Z., Shen, R., Jin, F., Wang, Y., ... & Liu, W. (2020, November). Deep-HOSeq: Deep higher order sequence fusion for multimodal sentiment analysis. In 2020 IEEE International Conference on Data Mining (ICDM) (pp. 561-570). IEEE.

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).

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).

Wang, X., Gao, L., Wang, P., Sun, X., & Liu, X. (2017). Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length. IEEE Transactions on Multimedia, 20(3), 634-644.

Wang, Y., Guan, L., & Venetsanopoulos, A. N. (2012). Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. IEEE Transactions on Multimedia, 14(3), 597-607.

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.

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

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

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).

YAO, Y., & JIANG, X. (2023). Video-based person re-identification method based on graph convolution network and self-attention graph pooling. Journal of Computer Applications, 43(3), 728.

Yolchuyeva, S., Németh, G., & Gyires-Tóth, B. (2020). Self-attention networks for intention detection. arXiv preprint arXiv:2006.1558

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

Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L. P. (2018, April). Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

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).

Zhao, B., Gong, M., & Li, X. (2022). Hierarchical multimodal transformer to summarize videos. Neurocomputing, 468, 360-369.

Zheng, Z., An, G., & Ruan, Q. (2018, August). Temporal pyramid pooling based relation network for action recognition. In 2018 14th IEEE International Conference on Signal Processing (ICSP) (pp. 644-647). IEEE.