**hello Dr**

**my name is ...**

**my supervisor is Dr suzan**

**my title is Image and Text Fusion for Intent Detection in Multimedia**

**I will introduce my research from these six parts.**

**first is background.**

读内容

**this is an example of intent detection using image and text**

**there are three major factors promote and affect the development of intent detection**

one is data.

读内容

The second is tool.

读内容

The third is Model

读内容

**Problem Statement**

读内容

**Research Objective & Question**

读内容

**Research Scope & Significance**

读内容

**Literature Review**

**for Intent Detection**

From the perspective of model structure, the related works can be divided into manual rules, statistical methods, deep learning.

From the perspective of model input, we can divide the related works into single-modal and multimodal.

the manual rules and statistical methods are both single-modal, most of deep learning methods are single-modal, and a small portion are multi-modal.

this is a brief list of references of intent detection.

For the single modality, the main use is text.

Different methods have different advantages and disadvantages.

for Manual Rules

easy to realize and fast operation speed, but high workload, poor stability.

in Statistical Learning Method

it can automatically learn pattern form big data, but need Manual feature extraction, so the efficiency is low.

in Deep Learning

such as Bert, can automatically extract feature and learn pattern, but this method only can be used for text.

some multimodal methods use different models to process different modalities, the stability and performance are relatively high, but the framework is More complex and need Higher computing and storage resources.

**references for Vision-Language Pre-training**

From the timeline, it can be seen that a series of representative multimodal large-scale pre-training models have appeared since 2019, mainly the Vision-Language model.

these multimodal pre-training models have been used in many practical tasks, such as Sentiment Classification、News Recommendation、Image Retrieval and so on, and has improved the performance in related fields compared with traditional methods.

**Research Methodology**

in this research, we use image and text to achieve intent detection, The main architecture is shown in Figure, which mainly includes three parts: feature representation, fusion, and classification. in the representation part, text and image feature extraction, alignment, and representation are automatically achieved using the CLIP pre-training model. in the fusion part, considering that different modalities contain different amounts of information, and have different contributions to intent detection, we design two-level cross-modality attention module to fuse character-level and global-level feature of image and text. Finally, the fused features are input into the classifier to achieve intent detection.

**CLIP**

CLIP is the abbreviation of Contrastive Language–Image Pre-training, this model is developed by OpenAI based on 400 million image-text data pairs. Adopting Vision Transformer (ViT) and the Text Transformer (Bert) structure and can process image and text simultaneously, highly efficient, flexible, and general. this model mainly consists of two parts: Text Encoder and Image Encoder to extract text and image feature respectively.

**Image Encoder**

The detail architecture of image Encoder is as shown in the left Figure, First, divide the image into patches and transform each patch to obtain input vectors. Then, add positional information to consider the sequence of inputs. finally, we get the global-level and character-level features as output.

**Text Encoder**

The detail architecture of text Encoder is as shown in the left Figure, it is similar with Image Encoder, because they both adopt classic transform structure.

**fusion strategies**

common multimodal fusion strategies can be divided into feature-level fusion, decision-level fusion, and hybrid fusion according to different fusion stage, in this research, we adopt feature-level fusion approach, which can better adjust the attention between modalities based on attention mechanism to achieve more effective information fusion.

**attention process.**

in the process of feature fusion, we adopt cross-modality attention mechanism to integrate information of different modalities, unlike Traditional vector splicing method, Cross-modality attention allows models to selectively focus on the most relevant parts of each modality. the process of the calculation is as shown in the image.

**datasets**

in this research, we used the latest publicly available intent detection dataset released by Tsinghua University in 2022. This dataset constructs a new hierarchical intent system, including two coarse-grained and 20 fine-grained intent categories. The statistics of the datasets are shown in the left Table, we split training, validation, and testing sets in 6:2:2. The detailed numbers of samples are shown in the right table.

**evaluation and detail**

In this experiment, Accuracy, precision (P), recall (R), and F1-score are used as the performance evaluation metrics of the model. Through observation of test results and based on prior knowledge, we set main parameters as shown in the table:

**Experiments on Intent Detection**

To verify the effectiveness of the proposed approach, we used three mainstream multimodal learning models and two mainstream single-modal learning models as baseline models to compare the results. Through the comparison with these five representative models, we can effectively evaluate the performance of the proposed method.

读内容

**Ablation Study**

This part is to verify the improvement of model performance by each module in this study, ablation experimental studies are carried out for different types of input data, feature representation module, and fusion module.

"-Text" means removing text. "-Vision" means removing image. "-CLIP" means removing the CLIP module. "-CAF" means remove the cross-attention feature fusion module.

读内容

**Influence of Encoders**

OpenAI provides two major types of pre-training models, one is based on RNN structure, and the other is based on transformer structure. To further explore the impact of different image encoders on performance, we perform a Contrastive experiment to evaluate four main image encoders.

读内容

**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.

读内容