In the field of human-computer interaction, human intention detection is a challenging problem and a key link to achieve 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 intent has become increasingly challenging. Traditional unimodal approaches, especially those relying solely on either textual or visual information, often fall short in 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 intent detection.

Compared with single-modal data such as images and text, multi-modal data can contain more information and can more accurately identify user intentions. At present, there are few studies on intent recognition based on image and text fusion, and they mainly consider how to integrate modal features in the feature fusion stage, while in the feature extraction stage, only a single-modality pre-trained model is used, which not only increases the complexity of the system , which increases the computational cost and may also limit the model’s comprehensive understanding of the global context, making it difficult to effectively capture the correlation information between modalities.

This research endeavors to design a new intention detection framework decomposing the multimodal learning problem into two equally important stages of representation and fusion, to explore the integration of image and text data to enhance the accuracy and robustness of intent detection in multimedia content. In the feature representation part, the CLIP multi-modal large-scale pre-training model is used to simultaneously extract text and image features, which simplifies system integration and saves computing resources while learning the associated information between modalities. In the feature fusion part, due to the different importance of text and pictures, an attention-based cross-modal fusion method is designed, which enables the model to dynamically adjust the attention to different modalities at each moment and capture important features, while reducing noise. In order to verify the effectiveness of the proposed model, this paper conducts experimental verification on the intent detection data set.

意图识别作为人工智能和计算机科学领域的重要研究方向，旨在使计算机系统能够理解和解释用户的真实意图，从而更智能地响应用户需求。近年来，随着多媒体技术的快速发展，用户发布信息的形式更加多样化，许多用户在发表文本信息时，通常配以相应的图片信息，来更加生动和直观地表达自己的真实意图，这种多媒体形式的信息更好地满足了用户在社交媒体上表达自己、获取信息和参与互动的需求，同时也给意图识别带来了新的挑战和机遇。

近年来，机器学习在处理图文等多种媒体形式上取得了显著的进展，尤其是深度学习技术的广泛应用，为意图检测提供了强大的工具，使得模型能够更好地从复杂的海量数据中学习和理解用户的真实目的。这些技术的应用推动了意图检测技术在各个领域的进步，为智能系统提供更为精确和智能的用户交互体验。但是，传统的意图识别方法通常局限于单一模态的数据分析，没有充分利用多媒体数据的丰富信息，在实际应用中，用户往往会同时使用图像和文本等多种形式进行交流，而单一模态的处理方法难以全面理解和识别用户的真实意图。

在数字化多媒体时代下，意图识别领域正面临着更加复杂和多样化的用户表达形式，而多模态数据融合作为一种前沿的研究方法，特别是图像与文本的融合，为解决这一问题提供了新的思路。图文融合相对于单个模态来说包含更多的信息，一方面，这些信息可以相互补充，在某个模态意图倾向表达不明确时，通过其他模态进行补充；另一方面也有可能出现虚假数据或相互矛盾的现象，如何正确利用和处理好各模态间的联系与矛盾，提高意图识别的准确性和鲁棒性，成为该领域的主要研究方向之一。

多媒体包括文字、图像、音频、视频等多种形式的信息表达方式。近年来，社交媒体平台的爆炸式增长为多媒体内容提供了广泛的传播平台，而多媒体形式的信息更好地满足了用户在社交媒体上表达自己、获取信息和参与互动的需求，但也带来了挑战，特别是在理解和预测这些数字空间中的用户意图方面。意图识别作为人机交互的核心技术之一，旨在从用户的输入中准确地识别出其意图，以实现更加智能化的服务，在多媒体技术中，通过意图识别了解用户的评论是真正的信息性、讽刺性、支持性还是批判性，对于完善推荐系统、内容审核和培养更健康的在线社区至关重要。而传统的意图识别方法存在一些潜在的薄弱点，例如往往只考虑了文本或图像等单一模态的数据，没有充分利用多媒体数据的优势。现有的基于图文信息融合的研究主要集中在情绪分析和内容分类上，但在解读用户行为背后的微妙意图方面仍存在重大差距。基于此，本文提出了基于图像和文本融合的意图识别方法，通过此研究，我们不仅可以验证图文信息融合技术在意图识别中的可行性和有效性，提高多媒体环境下意图识别的准确性，还可以使其更适应现实世界中多样化的用户交流场景。

传统的意图识别方法主要建立在人工规则的基础上，通过事先定义的规则集来推断用户的意图，这种方法虽然简单，但是需要手动整理规则集，覆盖范围有限且不够灵活。随着深度学习技术的发展，研究者们提出了许多基于深度学习的意图识别方法[13，14]，这些方法可以自动的从海量数据中学习到相关模式，相对于人工规则更加高效和准确，但是只使用单个模态（例如，文本模态）进行意图识别，往往不能充分利用信息的多样性，面对输入内容的多样性和复杂性难以有效完成任务。因此，多媒体环境下，如何有效的整合和利用多模态数据是当前该领域的主要研究问题之一。

Nowadays，various multi-modal technology related solutions and methodologies have emerged to address the above problem in intent detection, for instance:

In [1], a model is proposed to capture the complex meaning multiplication relationship between image and text in multimodal Instagram posts. While this model integrated text and images information to identify the intent, 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.

In [2] introduces a late-fusion approach for integration of the video signal with the captions signal for Intent Detection. Although it shows significant improvements with unimodal pretrained models, the HERO used in the article is a pretrained model only for video language, heterogeneity between modalities is not considered and its performance on image and text data is not yet known.

In [3] develops an adaptive multimodal fusion method based on an attention-based gated neural network, which can distinguish the contributions of different modalities. It designs complex strategies in feature fusion to reduce possible noise but use original pretrained models Bert and ResNet50 to extract text and image features respectively in feature extraction, which limits the model's comprehensive understanding of the global context.

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

**Research Questions**

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

**Research Objectives**

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

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 intent. For example, 【1】 compared the effects of bag-of-words model, TF-IDF and n-gram methods in short text intent analysis.【2】employ continuous bag-of-words coupled with support vector machines (SVM) to tackle the problem of intent classification.

With the development of deep learning, many intention recognition methods based on deep learning have been proposed 【3】, such as, 【4】 present a novel intent detection system which is based on a self-attention network and a Bi-LSTM，【5】propose a novel approach to intent recognition which involves combining transformer architecture with capsule networks。【6】developed an Intent 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。【7】 introduce intent 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【8】, multimodal technology can be used to analyze text and image information, identify users' emotional tendencies and expressions. In terms of humor detection【9】, 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 recognition. 【10】proposed a model to capture the complex meaning multiplication relationship between image and text in multimodal Instagram posts. 【11】proposed a late-fusion approach for integration of the video signal with the captions signal for Intent Detection.【12】introduced an adaptive multimodal fusion method based on an attention-based gated neural network, which can distinguish the contributions of different modalities.

随着自然语言领域预训练模型技术的逐渐成熟，多模态预训练模型也逐渐受到关注，一系列的视觉-语言预训练工作应运而生。视觉-语言预训练学习 VLP（Vision-and-Language Pre-training）［3］是指基于海量图像-文本对数据训练跨模态的通用表征，得到的预训练模型可以直接微调适配下游视觉-语言任务。根据编码方式的不同可以大概分为双塔编码和融合编码。

双塔编码主要关注图像和文本的各自模态编码的表征对齐，采用最简单点乘融合特征。目前热点模型如CLIP［2］和ALIGN［4］等，这类方法使用了对比学习进行预训练，采用余弦相似性来度量模态间的距离，并在不同领域展示了卓越的性能。近期 Meta AI 何凯明团队推出 FLIP［24］多模态预训练模型，融合了MAE［25］中的图像文本双掩码技术，可以在有限的时间内从更多的image-text数据集中学习，相对于CLIP有效提高了模型预训练的效率。

融合编码框架使用Transformer机制进行跨模态融合。ViLBERT［27］和LXMERT［28］提出使用三个不同的Transformer分别进行图像编码、文本编码和特征融合，再增加了融合阶段的网络深度后，混合编码模型框架在视觉-语言类下游任务中表现出优异的表征能力。但是这类算法受限于网络训练和推理速度，并未得到工业界的广泛应用。ViLT［32］针对推理速度问题进行了优化，通过简单化的网络设计，使用Transformer模型的encoder编码器来提取和处理视觉特征，而不是单独的计算机视觉模型来提取特征。实验显示，该方法在参数量和运行时间上都能明显降低，模型效果明显优于LXMERT等融合编码框架，但是和CLIP［2］双塔框架还是有一定差距。

After analyzing representative methods in the field of intent detection in recent years, it can be concluded that the research content of multimodal technology 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 feature extraction method is used. However, the development of multi-modal large-scale pre-trained models gives us new ideas for feature extraction and representation.

**方法论部分**

本研究旨在使用图像和文本多模态数据检测意图。主要架构如图2所示，主要包括三部分:特征提取和表示、多模态融合和分类。首先，在第一部分，使用CLIP多模态大规模预训练模型自动实现文本和图像特征提取、对齐和多模态表示。其次，在第二部分，考虑到不同模态是不同性质的数据，包含的信息量不一样，对意图检测的贡献程度也不同，设计了一种基于跨模态注意力机制的多模态特征融合方法。最后，将融合后的特征输入到分类器中实现意图识别。

特征提取

输入特征的质量对多模态意图识别模型的预测结果有着重要影响，早在机器学习时期，特征工程就决定了学习的上限。更好的特征意味着不需要复杂的模型也可以得出优秀的结果。随着深度学习神经网络的发展，特征提取的方式也有了很大变化，当前在多模态意图识别中，主要使用BERT和ResNet预训练模型提取文本和图像特征，BERT和ResNet通常是独立地训练的，这导致每个模型只能理解其特定模态的信息，限制了模型对于全局上下文的综合理解，而在图文任务中，模态之间的关联信息是非常关键的。多模态预训练模型相对于单一模态的预训练模型具有一些明显的优势，它采用对比学习等方法，在预训练阶段就能够学到模态之间的关联信息，因此能够同时处理多种模态的数据，提高了信息理解的能力，在多个图文任务中，超越了旧有的单一模态方案，表现出较强的迁移能力。而且一个单一的多模态预训练模型可以直接用于处理多模态任务，简化了系统的集成和管理。因此，本研究首次在意图识别领域使用多模态预训练模型CLIP，提取文本和图像的特征，实现多模态表示。

CLIP（Contrastive Language-Image Pre-training） 模型是由OpenAI基于4亿图文数据对开发的一种多模态预训练模型，在处理文本和图像的任务中表现出色，在很多任务表现上达到了目前最佳表现（SOTA）。它采用对比学习的方法进行预训练，它通过最大化相关图像和文本对之间的相似性，同时最小化不相关图文对之间的相似性，将图像和文本映射到一个共同的嵌入空间，这使得CLIP能够同时理解文本和图像。CLIP是在大规模的多模态数据集上进行预训练的，这种大规模的数据集有助于模型学习更通用的特征，也可以在特定任务上进行微调，使模型适应特定领域或应用，从而具备通用性和可迁移性，能够适应不同的应用场景。如下图所示，CLIP主要包括两部分：Text Encoder和Image Encoder，其中Text Encoder用来提取文本的特征，可以采用NLP中常见的masked self-attention Transformer；而Image Encoder用来提取图像的特征，可以采用最新提出的ViT-B/32 Transformer architecture。

ViT-B/32 Transformer architecture用于图像编码，它是一种基于Transformer的图像分类模型，其中ViT表示Vision Transformer，B表示基础版本，32表示图像被划分成了32×32的图像块。相比于传统的卷积神经网络（CNN），ViT模型采用了纯Transformer结构，将图像视为一系列patch序列进行处理，具有更好的全局感知能力和泛化性能。此外，ViT模型还具有可扩展性强的优点，可以通过增加模型的深度和宽度来提高性能。

Masked Self-Attention Transformer 是一种基于 Transformer 架构的深度学习模型，主要用于处理文本中的序列数据，具有强大的表达能力和泛化能力。它通过采用 Masked Self-Attention 机制，使模型能够关注输入序列中的不同部分，并根据上下文信息生成相应的输出。

融合

在基于图文融合的意图检测任务中，除了提取不同模态的特征，更重要的是如何将不同模态的特征进行融合，多模态特征融合是模型综合多个模态进行预测任务的重要过程。由于不同模态数据之间存在互补性和差异性，对结果的贡献度也不同，通过特征融合可以为模型预测提供更多有效信息，提升决策的准确率。

目前常见的多模态融合策略是特征级融合[31](Feature-level Fusion)、决策级融合[32](Decision-level Fusion)和混合融合[33](Hybrid Fusion)。决策级融合可以对不同模态使用适合的模型来训练，因此可以更好的提取单模态内部信息，具有较好的泛化性，但是每个模态使用不同的模型进行训练，不能很好的捕获不同模态之间的交互信息，容易忽略不同模态之间的关联性。混合融合方法设计灵活并且同时具有特征级融合与决策级融合的优点，但是该方法比较复杂，实现较为困难，容易造成过拟合问题，且适用于三种模态及以上的场景。

为了提取不同模态的深度特征，更好的融合不同模态之间的信息，本文采用特征级融合策略，基于跨模态注意力机制融合图像和文本特征，和简单的向量拼接方式不同，基于跨模态注意力机制的多模态融合是指在处理多模态数据时，采用注意力机制来动态调整模态之间的关注度，以实现更有效的信息融合。在多模态融合中，跨模态注意力机制允许模型能够在每个时刻动态地调整对不同模态的关注度，在捕获重要特征的同时排除噪音。通过这种方式，模型能够更好地理解多模态数据的整体结构，从而提高任务的性能。

分类

我们将融合层得到的向量输入到多层感知机中，对于本文中的意图识别来说，本质是一个多分类问题，可以使用SoftMax作为神经网络的最后一层，用于计算意图预测分数。SoftMax是一种激活函数，它可以将一个数值向量归一化为一个概率分布向量，且各个概率之和为1。

其中W和b分别表示线性层参数和偏置项。使用交叉熵作为损失函数，交叉熵（Cross Entropy）是Shannon信息论中一个重要概念，主要用于度量两个概率分布间的差异性信息。

n为总的意图的数量，y是样本标签的one-hot表示，y表示样本属于第i类的概率