Multi-modal Human Desire Understanding on Social Media Data

Name: Abdul Aziz ID: 18701032

Supervisor Name: Nihad Karim Chowdhury

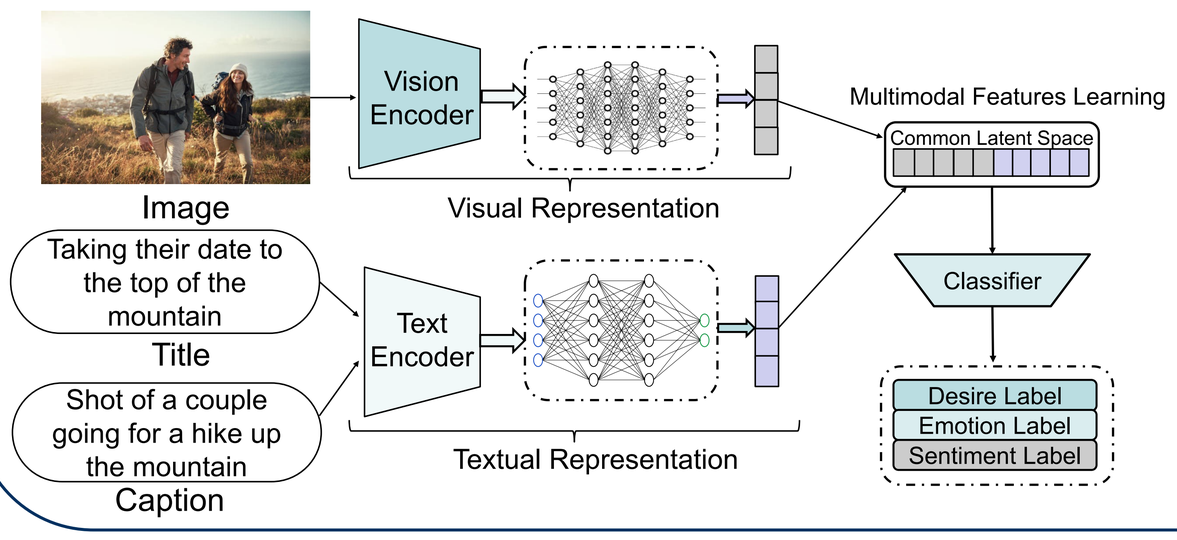
Department of Computer Science and Engineering, University of Chittagong, Bangladesh

Introduction

Problem Statement

Multimodal Human Desire Understanding:

Predicts the desire, emotion, and sentiment of image-text pair



Motivation

Multimodal human desire understanding provides a wide range of benefits:

- Effective human-computer interactions
- Recognize human emotional intelligence
- Understand interpersonal relationships
- Helps in decision making
- Helps to Evaluate human expressions
- Improve customer satisfaction and experience in e-commerce
- Helps to discover information about mass people's aspirations

Challenges

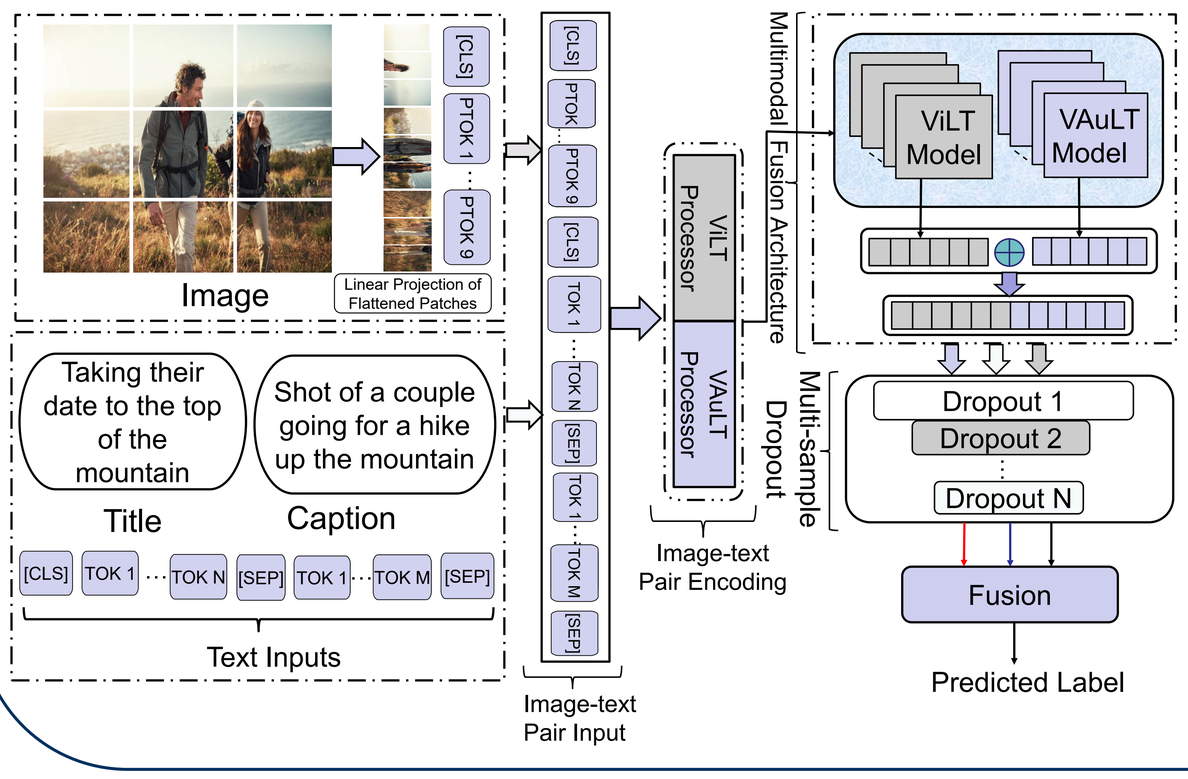
- Tightly coupled with sentiment analysis and emotion recognition tasks
- Combination of different types of modalities, including visual and textual content
- The diversity of cultures, countries, and languages are involved

Related Work

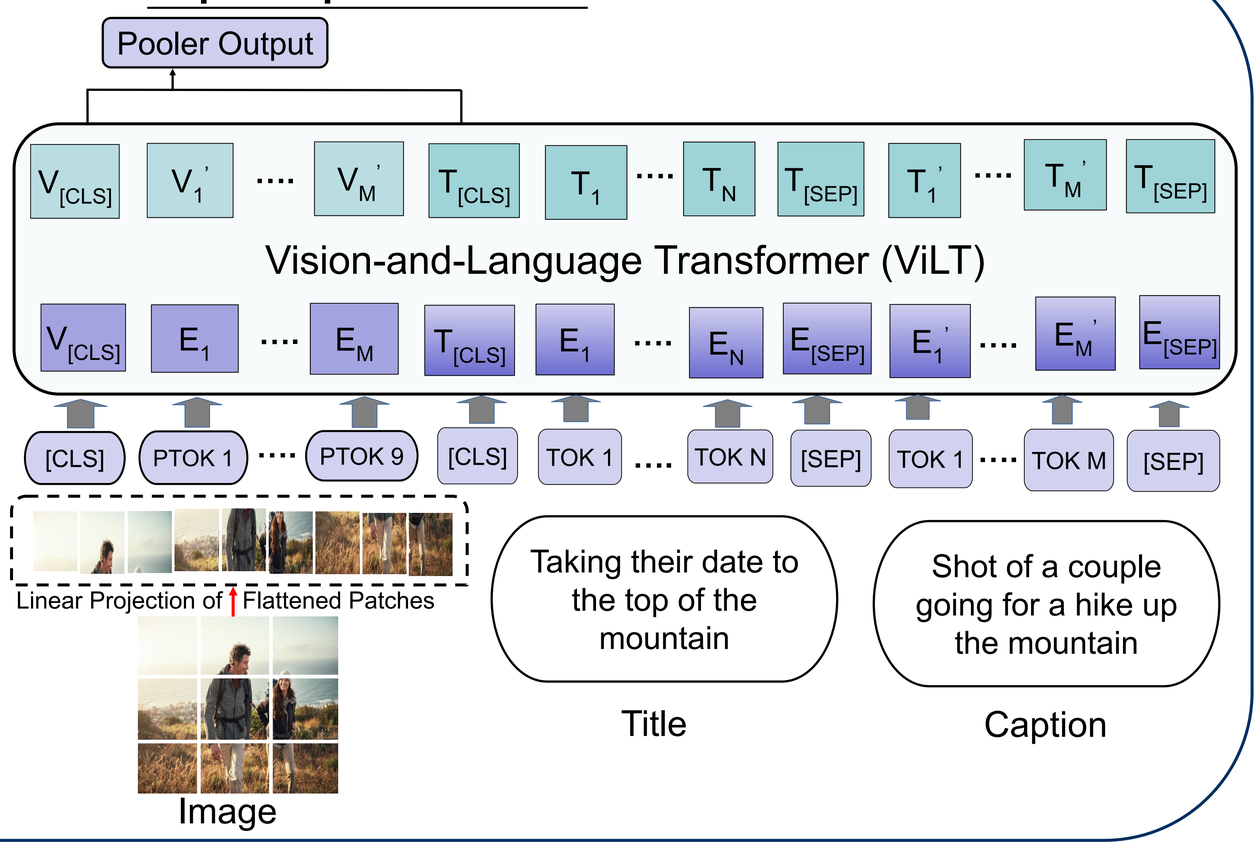
- Employed deep convolutional neural networks to extract features of visual and textual modalities focusing on the sentiment and emotion analysis tasks (Poria et al., 2016 IEEE ICDM Conf.)
- Proposed the first multi-modal dataset for human desire understanding and also provide various strong unimodal and multimodal baselines based on various visual and textual encoders (Jia et al., 2022 NAACL Conf.)
- Introduced a multimodal translation for sentiment analysis (MTSA) method leveraging text, visual and audio modalities where they improve the quality of visual and audio features by translating them into text features using BERT (Yang et al., 2022 Elsevier Neurocomputing Journal)
- M3GAT: Design a graph attention network-based method focusing on the sentiment and emotion analysis tasks (Zhang et al., 2023 ACM TOIS Journal)

Methodology

Proposed Desire Understanding Framework



Input Representation



Experiments and Evaluations

MSED: A multimodal desire understanding dataset, 9,190 text-image pairs gathered from Twitter, Getty Image, and Flickr. The MSED dataset is comprised of 6127 train, 1021 validation, and 2024 test instances

Overall Performance Across Three Tasks

Task	Precision	Recall	F1-Score
Sentiment Analysis	88.27	88.68	88.44
Emotion Analysis	84.39	84.64	84.26
Desire Analysis	84.23	82.01	83.11

Comparative Performance Analysis

Methods	Precision	Recall	F1-Score
Sentiment Analysis			
Multimodal Transformers	83.56	83.45	83.50
M3GAT	84.66	85.15	84.85
BERT+ResNet	85.83	85.79	85.81
MMTF-DES (Ours)	88.27	88.68	88.44

MMTF-DES: Multimodal Transformers Fusion for Desire, Emotion, and Sentiment Analysis

Comparative Performance Analysis

Methods	Precision	Recall	F1-Score
Emotion Analysis			
Multimodal Transformers	81.62	81.61	81.53
M3GAT	82.53	81.51	81.97
BERT+ResNet	83.54	81.51	82.42
MMTF-DES (Ours)	84.39	84.64	84.26
Desire Analysis			
Multimodal Transformers	81.42	80.20	80.92
BERT+ResNet	83.43	82.43	82.28
MMTF-DES (Ours)	84.23	82.01	83.11

Our proposed method, MMTF-DES outperforms existing approaches by 3% for sentiment analysis, 2.2% for emotion analysis, and approximately 1% for desire analysis in terms of macro F1 score

Improvement Over 7th Semester

- We proposed a multimodal transformers fusion model to predict the desire, emotion, and sentiment of the image-text pair
- Our method outperforms the SOTA methods across all tasks in terms of F1 score

Conclusion

- Addressed newly introduced human desire understanding task
- Proposed a unified multimodal transformer-based framework
- We provide a rigorous discussion of the human desire understanding task