ACM multimedia asia 2021

Hybrid Improvements in Multimodal Analysis for Deep Video Understanding

BeibeiZhang [1], Fan Yu,[1,2] Yaqun Fang,[1] Tongwei Ren[1,2,*]

[1] State Key Laboratory for Novel Software Technology, Nanjing University

[2] Shenzhen Research Institute of Nanjing University

NANJING UNIVERSITY 南京大学

MAGUS MediA recoGnition and UnderStanding

Introduction

- **Deep video understanding (DVU)**
  - requires systems to develop a deep analysis and understanding of long video.
  - use known information to reason about other, more hidden information, and to populate a knowledge graph (KG) with all acquired information.
- **HLVU** dataset
  - 14 videos
    - 10 for development
    - 4 for test
  - 1h/video in average
  - shot, entity name, entity type, screenshots

Training dataset:

1. Honey - Romance - 86 mins.
2. Let's bring back Sophie - Drama - 50 mins.
3. Nuclear Family - Drama - 28 mins.
4. Shooters - Drama - 41 mins.
5. Spiritual Contact The Movie - Fantasy - 66 mins.
6. Super Hero - Fantasy - 18 mins.
7. The Adventures of Huckleberry Finn - Adventure - 106 mins.
8. The Big Something - Comedy - 101 mins.
9. Time Expired - Comedy / Drama - 92 mins.
10. Valkaama - Adventure - 93 mins.

Testing dataset:

1- Bagman - Drama / Thriller - 107 mins.

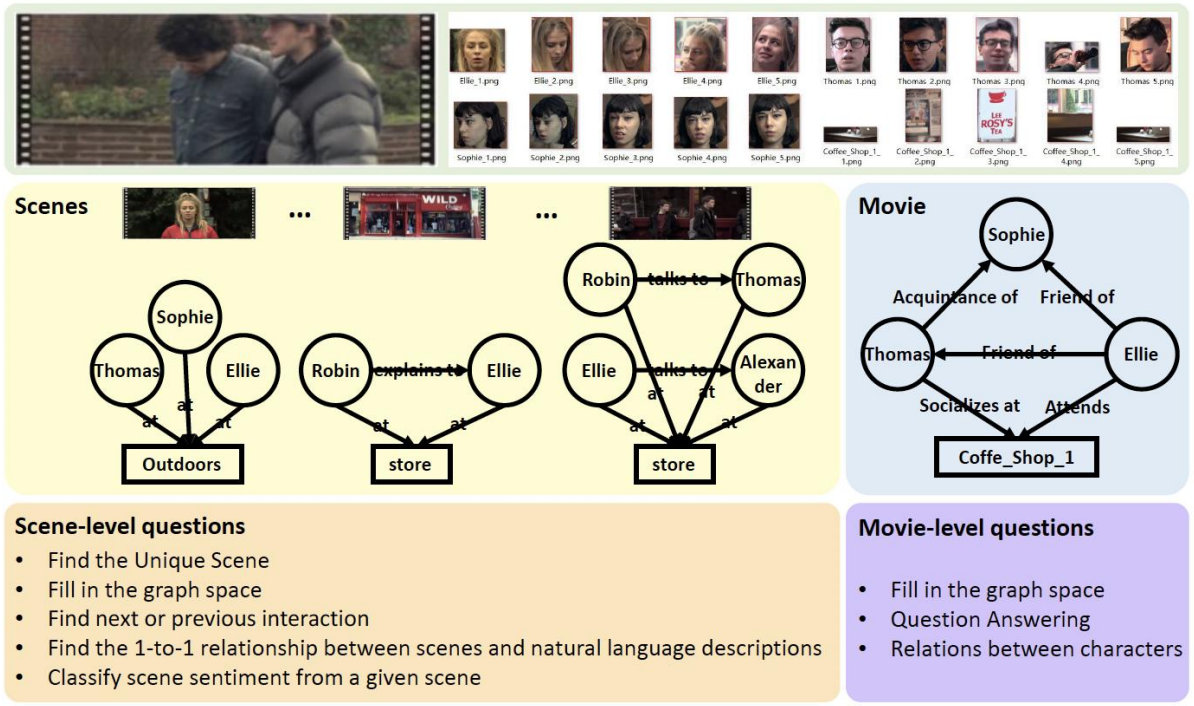
2- Manos - Horror - 73 mins.

3- Road to Bali - Comedy / Musical - 90 mins.

4- The Illusionist - Adventure / Drama - 109 mins.

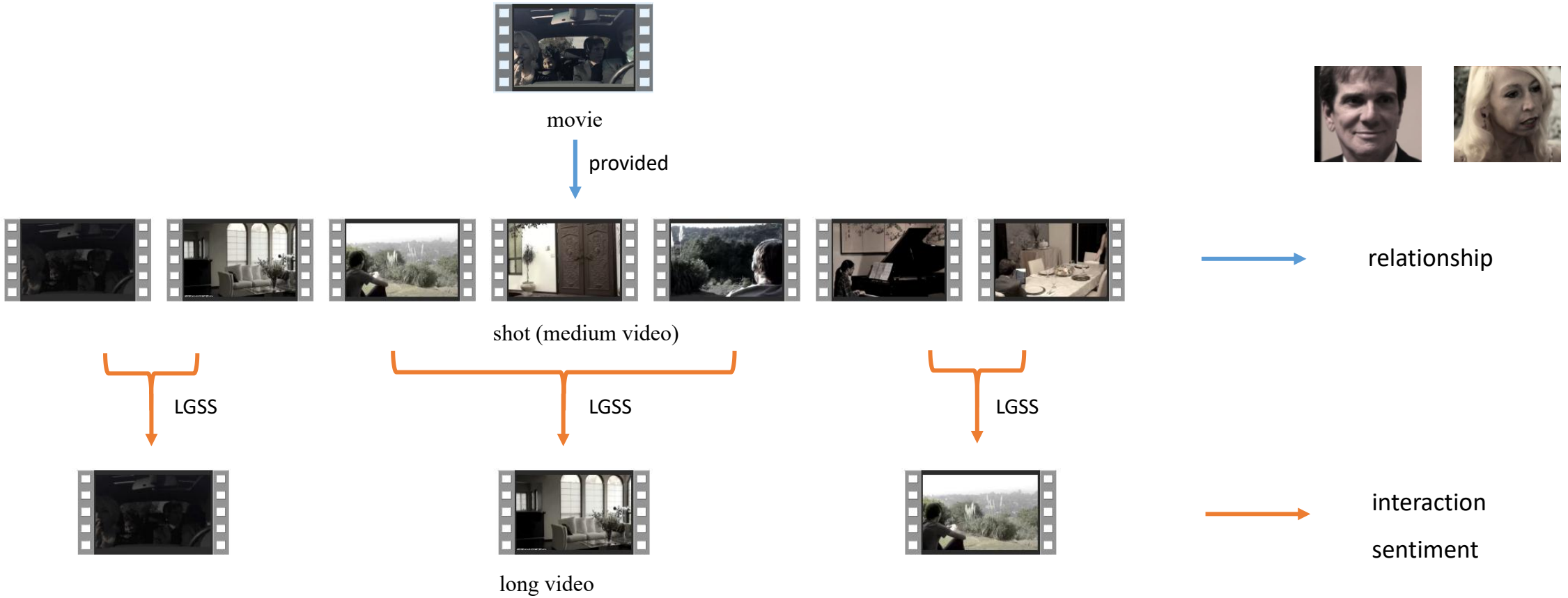
Task

- **scene-level**
  - Find the unique scene.
  - Fill in the graph space.
  - Find next interaction in scene X between person Y and person Z.
  - Find previous interaction in scene X between person Y and person Z.
  - Find the 1-to-1 relationship between scenes and natural language descriptions.
  - Classify scene sentiment from a given scene.
- **movie-level**
  - Find all possible paths question.
  - Fill in the part of graph question.
  - Multiple choice questions.



Pipeline

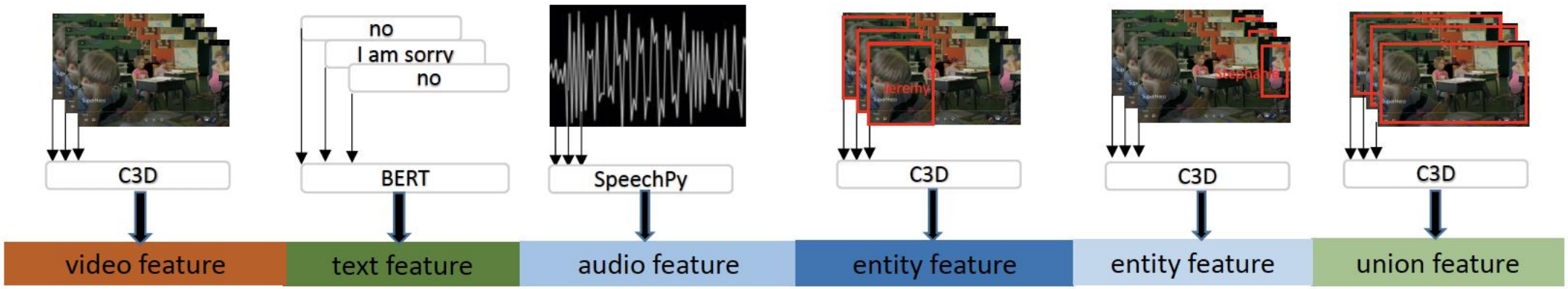
- Video segmentation



Pipeline

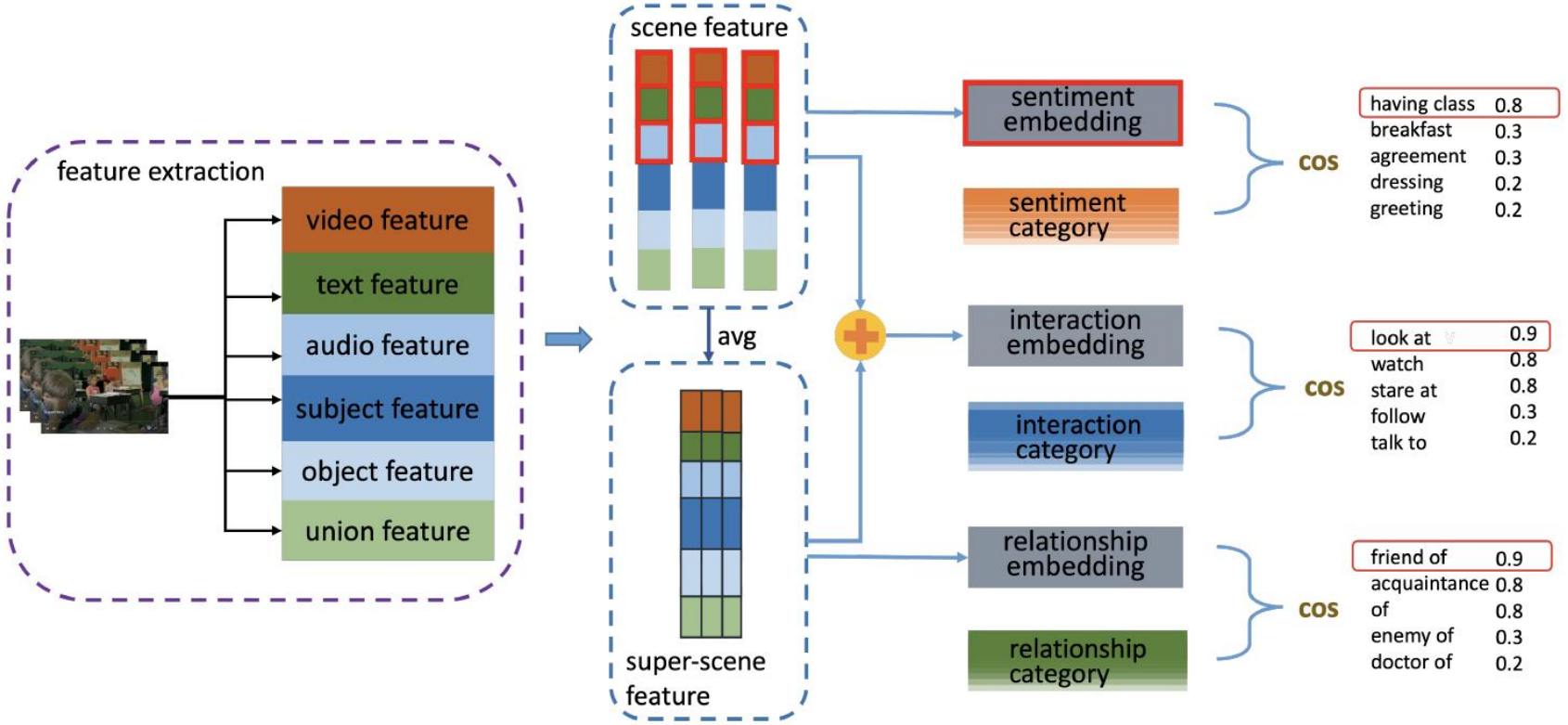
- **Video feature**
  - C3D
- **Audio feature**
  - MFCC, LMFE
- **Text feature**
  - BERT

- **Entity feature (subject, object, union)**
  - CenterTrack
  - InsightFace
  - C3D



Pipeline

- Joint learning architecture
  - relationship: average of medium video feature
  - interaction: medium video feature + average feature



Pipeline

- Low-shot, Zero-shot learning
- Joint learning

$$l = (1 - cos(\beta, \gamma))^2 + \frac{1}{n} \sum_{i \in U} (cos(\beta, \mu_i) + 1)^2$$

- l denotes loss
- denotes the feature of pair
- denotes the feature of the positive relationship
- denotes the set of negative relationships
- denotes the feature of relationship
- denotes the number of negative relationships

$$L = l_R + \frac{1}{n} \sum (l_I + l_S)$$

- denotes the total loss
- l_R denotes the loss of relationship
- $l_I$denotes the loss of interaction
- $l_S$denotes the loss of sentiment

Query answering

- **movie-level**
  - Find all possible paths question.
  - Fill in the part of graph question.
  - Multiple choice questions.
    - relationship knowledge graph
- **scene-level**
  - Find the unique scene.
  - Fill in the graph space.
    - interaction knowledge graph
  - Find next/previous interaction in scene X between person Y and person Z.
    - split medium video into shot videos
  - Find the 1-to-1 relationship between scenes and natural language descriptions
    - match with predicted interactions and sentiments.
  - Classify scene sentiment from a given scene.
    - sentiment model

Improvement

Ezra_1.png

SURF

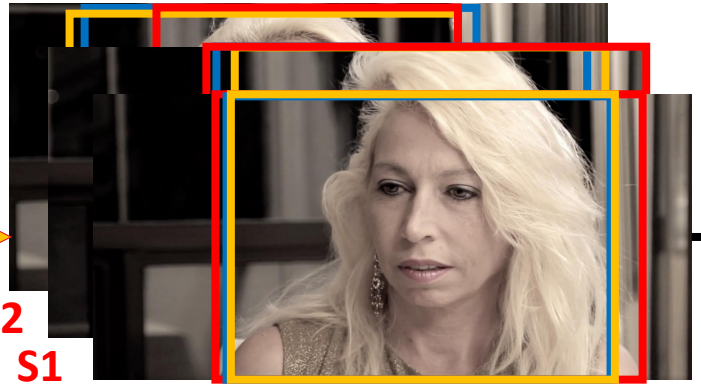
InsightFace

CenterTrack

TraDes

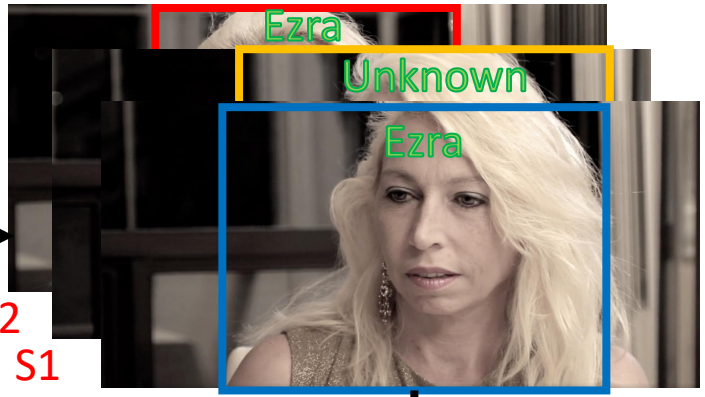
SiamMOT

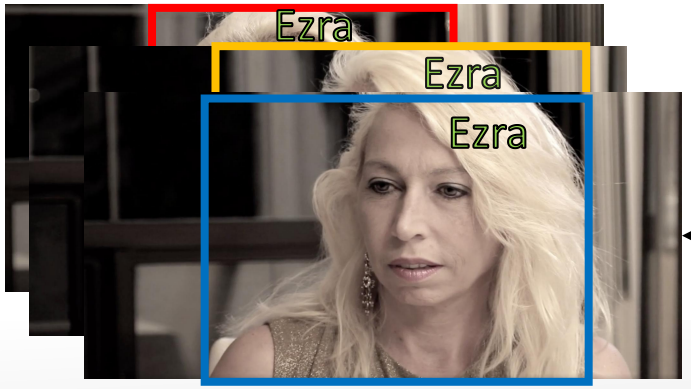
2 C2 T2 S2
1 C1 T1 S2
0 C1 T1 S1



merge

C2 T2 **S2**
C1 **T1** S2
C1 T1 S1





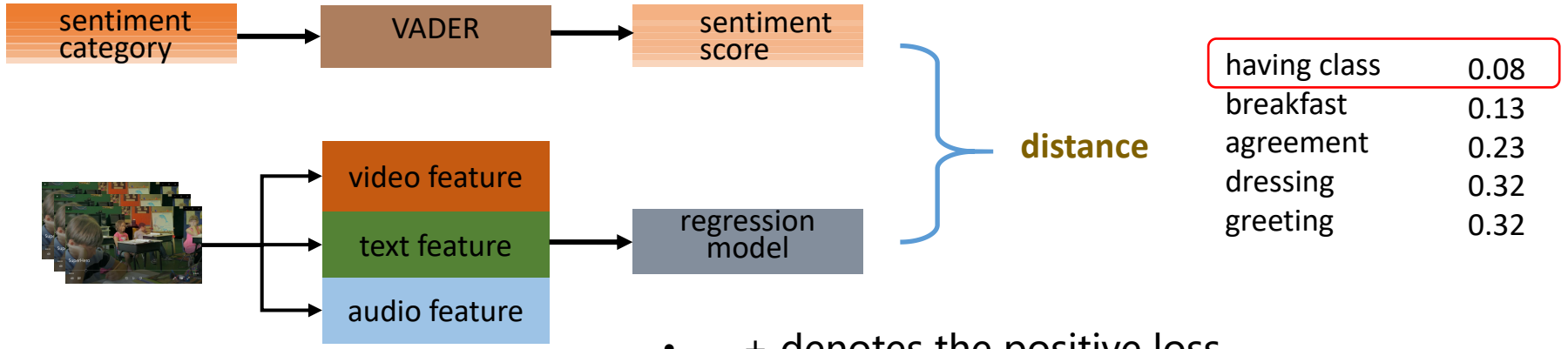
0: C1, Ezra
1: T1, **Unknown**
2: S2, Ezra

C1: Ezra 0:0 C2: Ezra 2:2
T1: Ezra 0:0 T3: Ezra 2:2
S1: Ezra 0:0 **S2:** Ezra 2:2

update

Improvement

- Sentiment score



- Sentiment distribution

$$l_+ = (1 - \cos(\beta, f_+))^2 \cdot d_+$$

$$l_- = (\cos(\beta, f_-) + 1)^2 \cdot d_-$$

- Experiment result

- + denotes the positive loss
- denotes the feature of scene,
- + denotes the feature of the positive sentiment
- + denotes the distribution ratio of positive sentiment
- − denotes the negative loss
- − denotes the feature of negative sentiment
- − denotes the distribution ratio of negative sentiment

	cos_sim	*emotion_score*	*senti_distribution*
Recall_k	27.5	**29.4**	**29.4**

Improvement

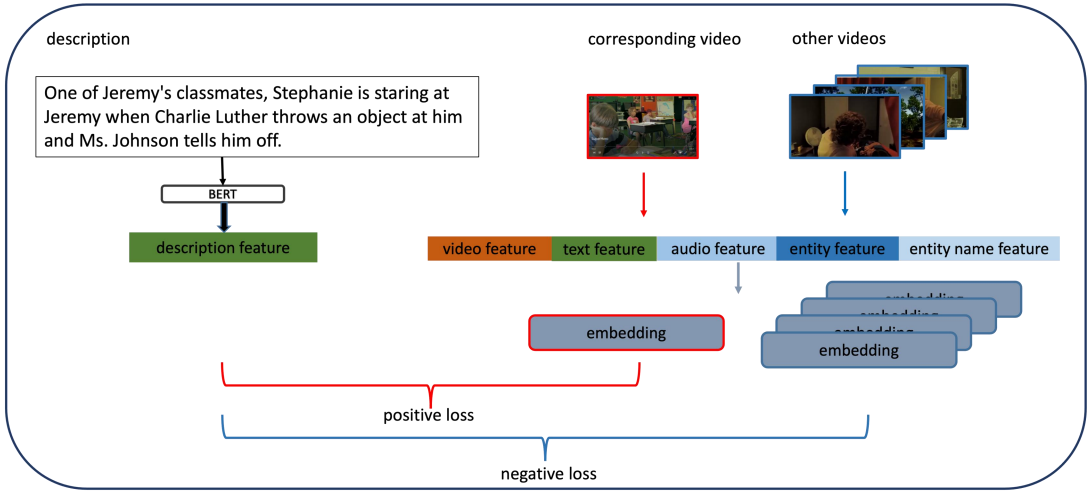
Model	C3D	I3D_{rgb}	I3D_{flow}	I3D_{avg}	I3D_{con}
Recall@50	28.8	29.1	**33.3**	32.8	29.6

- I3D represents rgb stream of I3D
- I3D represents optical flow stream of I3D
- I3D represents avaraging the outputs of rgb stream and optical flow stream
- I3D represents concatenating the features extracted by both of the streams

Improvement

Method	model	mat_I	mat_E	mat_{I+S}	mat_{E+O}	mat_{I+S+E}
Recall@k	4.9	4.9	12.1	5.3	**12.6**	6.8

- Model means description matching model
- mat means direct matching algorithm
- represents predicted interactions
- represents detected entities
- represents predicted sentiments
- represents detected objects.



THANK YOU


NANJING UNIVERSITY
1902
南京大學
NANJING UNIVERSITY


MAGUS
MediA recoGnition
and UnderStanding