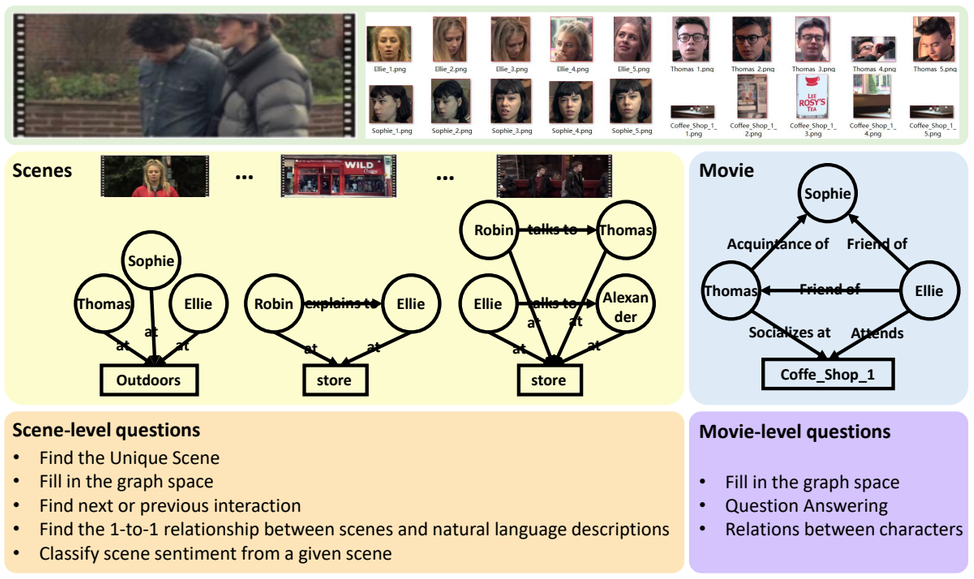
Hybrid Improvements in Multimodal Analysis for Deep Video Understanding

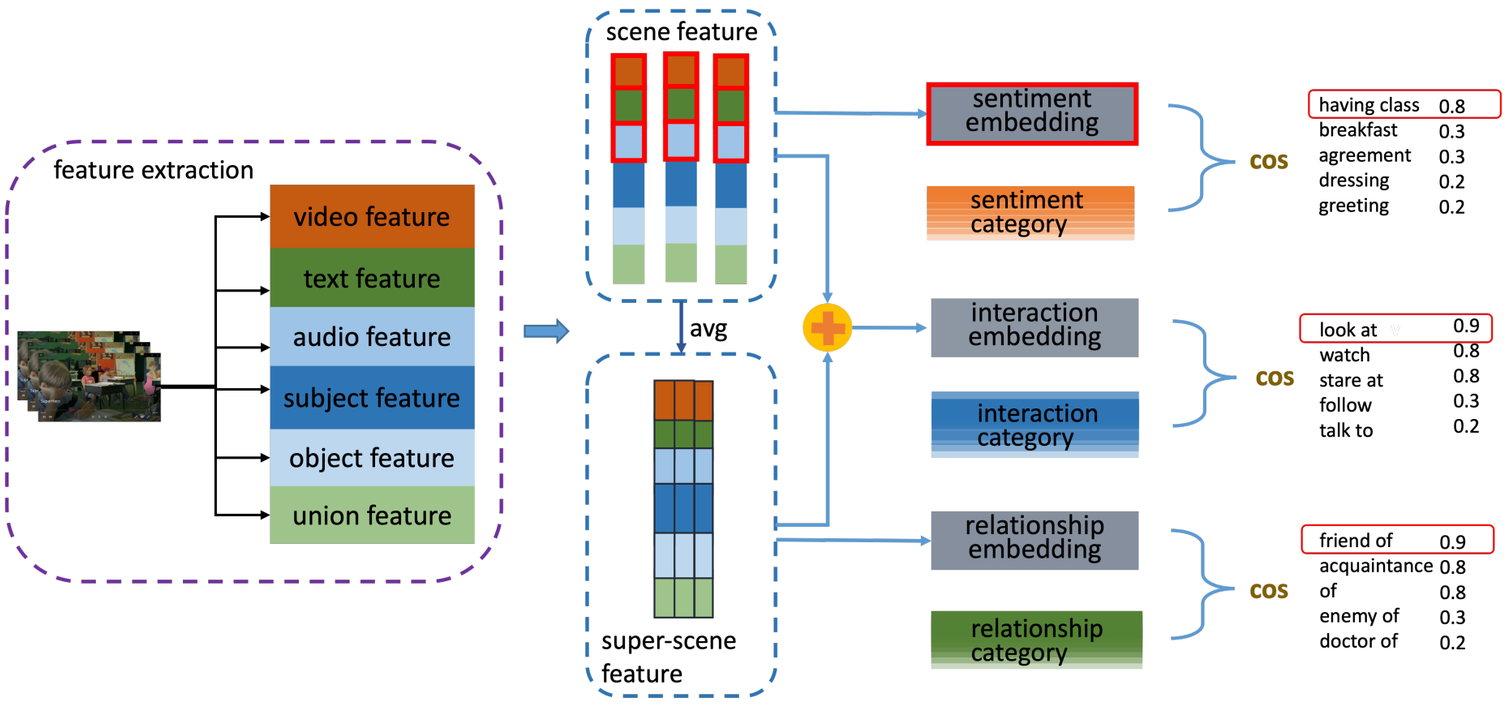
Beibei Zhang, Fan Yu, Yaqun Fang, Tongwei Ren, Gangshan Wu

Introduction

The Deep Video Understanding Challenge (DVU) is a task that focuses on comprehending long duration videos which involve many entities. Its main goal is to build relationship and interaction
knowledge graph between entities to answer relevant questions. In this paper, we improved the joint learning method which we previously proposed in many aspects, including few shot learning, optical flow feature, entity recognition, and video description matching. We verified the effectiveness of these measures through experiments.



Pipeline



The joint learning method we proposed extracts visual, audio and subtitle features from each scene and concatenates them together as scene feature. Then, the features of all scenes that comprise a super-scene are averaged as super-scene features and concatenated to each scene feature. Finally, we use super-scene features to predict relationships and scene features to predict interactions between entities, respectively.

Relationship and interaction prediction branches are trained together, which reflects the effects which physical relationships and interactions exert on each other.

As DVU challenge also provides sentiment data, we therefore add sentiment prediction branch to the joint learning framework.

Improvements

Few shot learning

- **Emotion_score**: predicting emotion score of scene with a regression model
- **Sentiment distribution:** multiplying the distribution of sentiment categories to loss as a cost

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

Optical flow feature

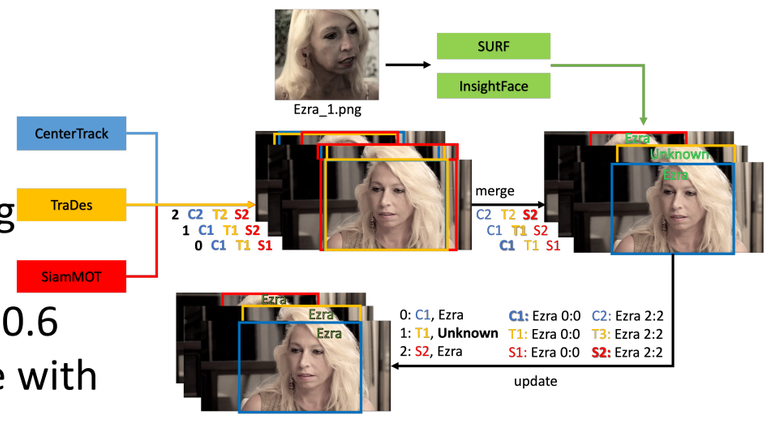
We introduced optical flow as a new modality, and used I3D to extract visual features.

- **I3D*avg***: avaraging the outputs of rgb stream and optical flow stream of I3D
- **I3D*con***: presents concatenating the features extracted by both of the streams

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

Entity recognition

- combining the results of CenterTrack, TraDes and SiamMOT
- merging the inter-covering bounding boxes when the IoU of them is more than 0.6
- recognizing character face with RetinaFace and SURF



Video description matching

We designed a video description matching model on the basis of multimodal features, and also experimented on direct matching with different elements.

- I represents predicted interactions
- E represents detected entities
- S represents predicted sentiments
- O represents detected objects.

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

南京大學 NANJING UNIVERSITY
177
zhangbb@smail.nju.edu.cn
MAGUS MediA recoGnition and UnderStanding