Hybrid Improvements in Multimodal Analysis for Deep Video Understanding

Beibei Zhang
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
zhangbb@smail.nju.edu.cn

Fan Yu
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
Shenzhen Research Institute of Nanjing University
Shenzhen, China
yf@smail.nju.edu.cn

Yaqun Fang
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
fanged@smail.nju.edu.cn

Tongwei Ren*
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
Shenzhen Research Institute of Nanjing University
Shenzhen, China
rentw@nju.edu.cn

Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
gswu@nju.edu.cn

ABSTRACT

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.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Deep video understanding; relationship analysis; interaction analysis; few shot learning

1 INTRODUCTION

Deep video understanding (DVU) is a task aiming to understand content of long duration videos on the basis of multimodal analysis [2]. The details of the task are shown in Figure 1. However, since videos are unstructured, it is hard to understand their content. Firstly, entity recognition is the basis and the prime challenge of video analysis. A long duration video may contain many entities, and the appearance of each entity may be different in different scenes. Secondly, interaction between entities may be complex and relation between entities may change over time. In terms of answering questions for video analysis, problems also need to solved.

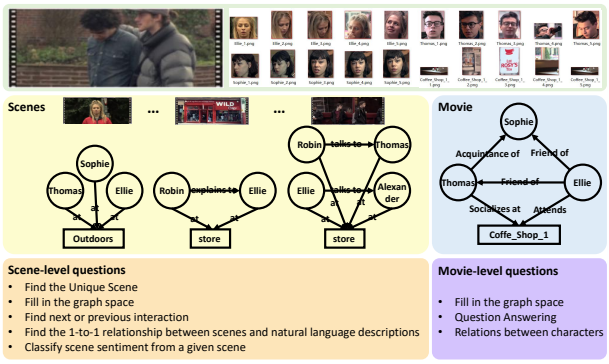

Figure 1: The description of deep video understanding task.

Our previous work [16] proposed a solution for answering the movie-level questions and later we further proposed an extended method [17] for answering both movie-level questions and scene-level questions by joint learning. As the basis, the designated scenes are divided into several shots, i.e., video clips without montage, and the designated scenes are merged into super-scenes, each of which contains a complete film section occurring at a certain place. The entities are recognized by tracking their traces in shots, scenes and super-scenes. Video visual features, audio features and text features are fused in the three modules. In the scene sentiment classification module, the video visual feature, audio feature and text feature of a designated scene are fused to generate the sentiment embedding for scene sentiment prediction.

*Corresponding author.

In the super-scene video relationship recognition module, super-scene relationship embeddings are generated from the fused video visual features, audio features and text features at scene level, along with visual features of two entities and visual union features of the entity pair at scene level by mean pooling. In the scene interaction recognition module, scene video visual feature, scene audio feature, scene text feature, scene visual features of two entities and scene visual feature of an entity-pair are fused together to generate the interaction embedding for predicting the interaction between two entities. All the three modules are trained jointly as an end-to-end model. According to the scene interaction embeddings and movie relationship embeddings, scene interaction knowledge graph and movie relation knowledge graph can be built by computing distances between candidate interactions' embeddings and candidate relations' embeddings, and most questions can be answered. To answer the questions about scene sentiment classification, the distances between scene sentiment embeddings and embeddings of candidate answers need to be computed. As for matching scenes and descriptions, the description is chosen according to the key words about interactions, relations between mentioned entities.

From the analysis on tasks and existed methods, we find some challenges. 1) It is difficult to track entities accurately in long video. 2) It is hard to perform well in sentiment classification. For the first problem, we suppose that the previous entity recognition sub-module and the feature extraction mechanism are not robust enough for different scenes in different movies. For the second problem, we consider that it is related to there are only a few samples of a sentiment category for training the scene sentiment classification module.

Thus, we propose four main improvements in this paper: 1) Blended object tracking is used for entity extraction. 2) Optical flow features are added for entity feature fusion. 3) Few shot learning is introduced for sentiment analysis. 4) Descriptions are matched with different strategies.

2 PRELIMINARY

Entity tracking and recognition in video. With some images of entities are provided, face recognition is an effective approach to tracking people in videos. RetinaFace [3] provides a valid solution for face box prediction and 2D facial landmark localisation. However, people in videos, especially in movies, are sometimes shot in close-up view and the whole figures may also appear. Thus, face detection results along with object tracking results are necessary to provide a useful solution to complete person tracking and recognition. CenterTrack [18] containing a one-stage object detector is a real-time tracking-by-detection method. TraDes [15] follows the joint detection and tracking paradigm and exploits the motion clue from tracking to enhance detection. SiamMOT [9] is a region-based siamese network for multi-object tracking, which estimates motion between two frames and the detected instances are associated.

Video visual feature extraction. Abundant networks have been proposed to extract image features, *e.g.*, AlexNet [7], VGG [11], ResNet [4], etc. Though videos are composed by multiple images, videos contain additional features: motion between frames. Simonyan *et al.* [10], propose a two-stream convolution network to incorporate spatial and temporal features. Tran *et al.* [13], use deep 3D convolutional networks for spatiotemporal feature learning. Carreira *et al.* [1], introduce a new two-stream inflated 3D convolutional network, which expands filters and pooling kernels of 2D convolution networks into 3D convolution networks.

Few-shot learning. Few-shot learning requires networks to classify samples into several classes where each class is only described with few examples. The key task for few-shot learning is to learn a function for similarity computation. Koch *et al.* [6], propose a siamese neural network to rank similarity between inputs. Santoro *et al.* [8], propose a memory-augmented neural network to rapidly assimilate new data and make accurate predictions with a few samples. Vinyals *et al.* [14], learn a network to map a small labelled support set and an unlabelled example to its labels. Snell *et al.* [12], propose prototypical networks that compute distances to prototype representations of each class to learn a metric space for classification.

3 OUR METHOD

As shown in Figure 2, the joint learning method we proposed before 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.

In this paper, we show improvements of the previous method in different aspects. As the number of sentiment training samples are small, we designed new methods for few shot learning. We also introduce optical flow features to help interaction prediction. In order to increase the accuracy of entity detection, we improved the MOT method. To answer the questions about scene description, we design a multi-modal model to match description and scene.

3.1 Few shot learning

Sentiment in the dataset is characterized by many categories, few samples and serious data bias. In our previous research, we predicted sentiment by embedding scene features and then calculating cosine similarity with sentiment category features. It has been a few shot learning method to compare feature similarity instead of directly predicting categories. In this paper, in order to carry out few shot learning for sentiment, we tried two new schemes.

One is to assign emotion score to each sentiment category through VADER [5] , and then use a regression model to predict emotion score using scene features. Finally, we calculated the distance between predicted scores and the emotion score of sentiment catagories.

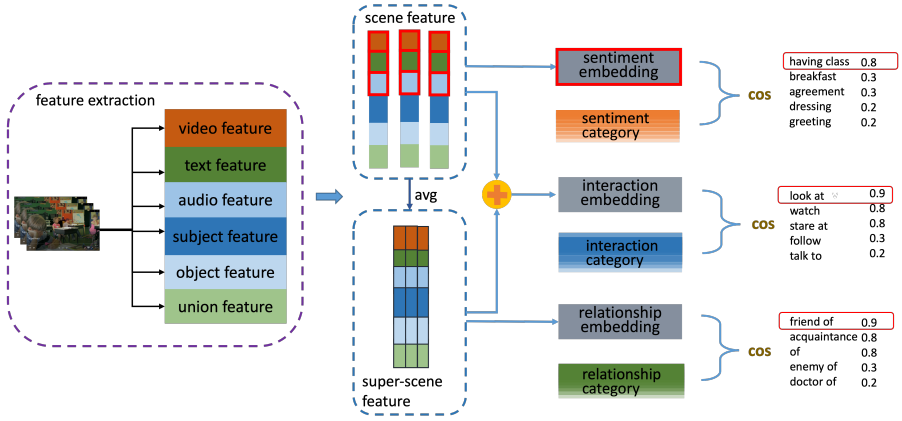

Figure 2: The joint learning method extracts mutiple modal features and concatenate them as scene feature. Then we average scene features as super-scene feature and concatenate to scene feature again. We embedded features and calculate cosine similarities between them and category features to get the final result. Relationship, interaction and sentiment prediction branches are trained together.

The other is to calculate the distribution of sentiment categories in the training set, and then multiply it to loss as a cost, as follows:

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

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

where l_+ denotes the positive loss, β denotes the feature of scene, f_+ denotes the feature of the positive sentiment, d_+ denotes the distribution ratio of positive sentiment, l_- denotes the negative loss, f_- denotes the feature of negative sentiment, , d_- denotes the distribution ratio of negative sentiment.

In this way, negative loss of the sentiments of small-sample categories increases, while for large-sample categories, positive loss of the sentiments increases.

3.2 Optical flow feature

As mentioned in I3D [1], I3D performs better than C3D in action recognition task due to the introduction of optical flow. According to the correlation between action recognition task and interaction prediction task, we used I3D to extract visual features. Due to limited developing data, We used I3D models pretrained by ImageNet, which outperforms Charades and Kinetics.

3.3 Entity recognition

Entities are divided into person, location and concept. In this paper, location is still recognized according to SURF and scene segmentation as in the previous method. Considering that in the previous work we only use CenterTrack [18] for person tracking with some traces omitted, we combine the results of CenterTrack, TraDes [15] and SiamMOT [9] to track and recognize people. Since the three methods could track the same person repeatedly, we compute the intersection of union (IoU) between each two bounding boxes at each frame and two bounding boxes need to be merged when the IoU of them is no more less than 0.6. When merging the inter-covering bounding boxes, we keep the smaller bounding box and tag the other bounding box with the id of the former one. However, people cannot be recognized only with tracking results. Similar to our previous method, RetinaFace [3] is used to recognize faces of people and SURF is used for matching templates of people. When there is no results of SURF and face recognition at a certain frame, the corresponding tracking bounding boxes are mapped to an "Unknown" person. Otherwise, a tracking bounding box is mapped to a person whose face/template has the largest covering over the tracking bounding box. If a person is mapped with multiple tracking bounding boxes at a frame, the bounding box with the largest covering over the face/template is kept and the information about person name, start frame index and end frame index of each tracking id is recorded. After the face recognition results and the SURF results at all frames are processed, we re-map the tracking results that were mapped before to the "Unknown" person to a new person, i.e., the one mapped with a tracking bounding box that have the same tracking id with the current tracking bounding box during a frame window.

3.4 Video description matching

We design a video description matching model on the basis of multi-modal features. We use BERT to export description features, and concatenate visual, audio, subtitle, character and character name features of each scene as video features. One scene corresponding to the description is positive sample, and other scenes are negative samples. Similarities between description and each scene are the outputs of the model.

3.5 Query Answering

Movie-level. We answer the following three types of queries: 1) To find all the possible paths, we construct a movie graph according to the relationship between entities, using entities as nodes and relations as edges. We use a depth-first search method to find all the possible paths between the source entity and the target entity, and use the confidence score of the relationship as the threshold for pruning. 2) To fill in the graph space, we traverse the edges of movie graph, obtain candidate edges that match the queries, and sort them according to the scores generated by our method. 3) We traverse all the choices, check whether the movie graph has an

edge that satisfies the conditions, and choose the best match as the answer.

Scene-level. We answer the following five types of queries: 1) To find the unique scene, we match the given interactions with scene-level entity-interaction graphs and select the scene with the highest matching score. 2) To fill in the graph space, we traverse the edges of the entity-relation graph, match the interactions where the predicate and the object are exactly the same, add the subject to the candidate list, which is sorted by the number of occurrences of the subject. 3) To find the next or previous interaction, we divide each scene into smaller acts in chronological order, generate the interaction sequence between two given entities, and then judge the next or previous interaction. 4) To match scene with natural language description,we use WordNet to implement word lemmatization on descriptions, and match the entities, objects, interactions and sentiments contained in the given scenes. 5) To classify scene sentiment, we choose the sentiment with the highest predicted score as the answer.

4 EXPERIMENTS

4.1 Dataset and Experimental Settings

All the experiments are conducted with E5-2680 v4 2.40GHz 14 cores CPU, 64GB memory and one GeForce RTX 3090 GPU, on the HLVU dataset [2].

4.2 Few shot learning

In experiments of few shot learning, we evaluate the performance of sentiment prediction using metric *Recall@k*, where k means the number of ground truth sentiments. According to table 1, compared with the original method of directly calculating feature similarity, the regression model for predicting emotion score and the method of introducing sentiment category distribution to loss both have positive influence. It proves the effectiveness of our improved measures to few shot learning. It is worth noting that data bias still has some influence after we examining the predicted results.

Table 1: Experiments on few shot learning methods.

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

4.3 Optical flow feature

As shown in Table 2, we compared the training results using C3D and I3D by calculating *Recall*@50 because interactions will be sorted by confidence scores while answering queries, and found that using optical flow feature alone to predict interaction was the best. The fusion of optical flow stream and rgb stream of I3D will make the result worse. We think it is on the one hand due to the difference between action recognition task and interaction prediction task. In addition, our few shot learning method does not directly predict categories, but compares feature similarity.

Table 2: Experiments on video models, where I3D_{avg} represents avaraging the outputs of rgb stream and optical flow stream, I3D_{con} represents 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

4.4 Video description matching

As shown in Table 3, we compared the results of the multi-modal video description matching model which is referred to in section 3.4 and the direct matching algorithm by caculating *Recall@k*. We found that the matching algorithm performs better, which means the complexity of the task requires us to bring in more information like detected objects of the video. We used object tracking results in the help of Centertrack [18] while matching, and the results show that objects information is necessary.

Table 3: Experiments on description matching, where model means our description matching model, mat means direct matching algorithm, $_I$ represents matching descriptions with interactions, $_E$ represents detected entities, $_S$ represents sentiments, $_O$ represents 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

5 CONCLUSIONS

In this paper, we have improved our previous method [17] in many aspects, such as few shot learning, optical flow feature, mot and video description matching. However, due to the characteristics of small sample and serious bias of data, how to design effective few shot learning methods is still a big challenge for DVU.

ACKNOWLEDGMENTS

This work is supported by National Science Foundation of China (62072232), Natural Science Foundation of Jiangsu Province (BK20191248), Science, Technology and Innovation Commission of Shenzhen Municipality (JCYJ20180307151516166), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[2] Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. HLVU: A New Challenge to Test Deep Understanding of Movies the Way Humans do. In *International Conference on Multimedia Retrieval*. 355–361.

[3] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5203–5212.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[5] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.

[6] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, Vol. 2. Lille.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

[8] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065* (2016).

[9] Bing Shuai, Andrew Berneshawi, Xinyu Li, Davide Modolo, and Joseph Tighe. 2021. SiamMOT: Siamese Multi-Object Tracking. In *CVPR*.

[10] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199* (2014).

[11] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[12] Jake Snell, Kevin Swersky, and Richard S Zemel. 2017. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175* (2017).

[13] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *IEEE International Conference on Computer Vision*. 4489–4497.

[14] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems* 29 (2016), 3630–3638.

[15] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. 2021. Track to Detect and Segment: An Online Multi-Object Tracker. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Fan Yu, DanDan Wang, Beibei Zhang, and Tongwei Ren. 2020. Deep Relationship Analysis in Video with Multimodal Feature Fusion. In *ACM International Conference on Multimedia*. 4640–4644.

[17] Beibei Zhang, Fan Yu, Yaqun Fang, Tongwei Ren, and Gangshan Wu. 2021. Joint Learning for Relationship and Interaction Analysis in Video with Multimodal Feature Fusion. In *ACM International Conference on Multimedia*.

[18] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2020. Tracking objects as points. In *European Conference on Computer Vision*. 474–490.