Multimodal Analysis for Deep Video Understanding with Video Language Transformer

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

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

Tongwei Ren*
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, 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 (DVUC) is aimed to use multiple modality information to build high-level understanding of video, involving tasks such as relationship recognition and interaction detection. In this paper, we use a joint learning framework to simultaneously predict multiple tasks with visual, text, audio and pose features. In addition, to answer the queries of DVUC, we design multiple answering strategies and use video language transformer which learns cross-modal information for matching videos with text choices. The final DVUC result shows that our method ranks first for group one of movie-level queries, and ranks third for both of group one and group two of scene-level queries.

CCS CONCEPTS

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

KEYWORDS

Deep video understanding; Relationship analysis; Interaction analysis; Sentiment analysis;

1 INTRODUCTION

As deep learning has developed to this day, many specific visual recognition tasks have been researched for a long time. There are more and higher requirements on the understanding of videos. Besides, multimedia data, such as video, audio and text, are more and more popular. The Deep Video Understanding Challenge (DVUC) combines the requirements of video understanding and multimedia data, which requires using rich multimodal information to obtain more comprehensive inferences about videos, and automatically build richer, higher-level understanding [5]. Its main task is to predict the relationships and interactions between entities based on the provided videos, clips and entity screenshots. In addition, we need to find the relations between videos and sentiments or descriptions. Finally we can use the prediction results to generate the knowledge graph of the video to answer queries as shown in Figure 1. This challenge is difficult in three ways:

1) The videos last too long, each of which are about ninety minutes, and varies frequently over time. 2) Too many sub-tasks are involved, such as multiple object tracking, relationship prediction and interaction prediction. 3) The amount of data is limited, but the number of categories is large, resulting in few shot or even zero shot problem.

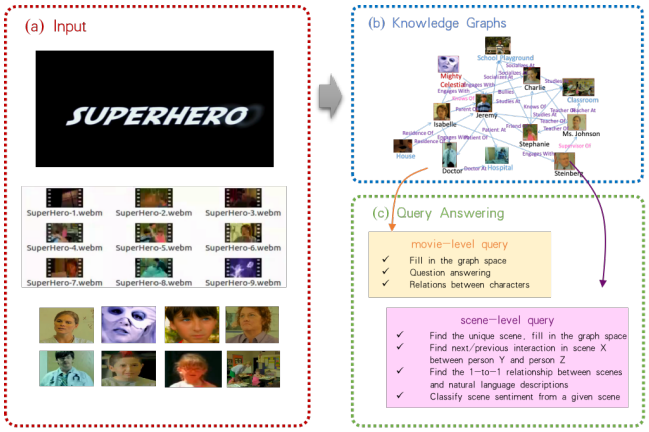

Figure 1: The illustration of deep video understanding challenge. (a) The input of DVUC are videos, clips and entity screen shots. (b) We need to use the input to generate knowledge graph of the video. (c) Finally, we can answer the queries of DVUC on the basis of the knowledge graph.

In our previous work[20, 22, 23], we have combined features such as video, audio and text to input a joint learning framework for multi-task prediction including relationship, interaction and sentiment recognition. In this paper, considering that actions are crucial for identifying human interactions and can indirectly affect the prediction of human relations, we introduce pose features and calculate pose interaction feature between humans as a new feature to facilitate interaction prediction.

The Deep Video Understanding Challenge is set up in the form of query answering, which can be viewed as a cross-domain problem. We introduce a video language transformer (VLM)[18] , which integrates cross-modal knowledge, to provide a better understanding of the questions. In addition, as the DVUC questions are very complex, we experiment on different matching strategies to improve the accuracy of the answers.

*Corresponding author.

In general, this paper has the following contributions:

1) Pose feature is introduced as a new feature into the multi-modal joint learning framework for multi-task prediction. It can represent actions well to improve the performance of interaction prediction.

2) VLM which is a video language transformer that learns cross domain knowledge is applied for understanding cross-domain questions more accurately.

3) Different matching strategies are designed to improve the accuracy of query answering.

As the final leaderboard shows, DVUC splits the queries into two groups of movie level and two groups of scene level, and our method finally ranks first for group one of movie-level queries and rank third for both of group one and group two of scene-level queries, which shows that our method is effective.

2 PRELIMINARY

Video Interaction Recognition With Pose Estimate. Video interaction detection aims to detect the interaction between people, which is a branch of Video visual relation detection[14]. Kukleva et al. focus on the interactions and social relationships between characters in a movie and predict interactions and social relationships with visual and language cues by joint learning[8]. Zhang et al. propose a solution to the previous DVUC task with multimodal feature fusion[23].

The human pose estimation[2] is the basis for many computer vision tasks,e.g., action recognition, behavior recognition, and unmanned driving,which aims to predict the position and relationship of human skeleton points, and there are some work aimed to extracting human 2d skeleton points from videos[4, 12, 15].

Recently, due to the great progress of pose estimation, many relation recognition tasks try to fuse human keypoint features for interaction recognition. Yun et al.[21] find that the geometric relational features based on the distance between all pairs of joints outperforms other feature choices. Luvizon et al.[11] propose a multi-task framework for jointly 2D and 3D pose estimation from human action recognition in video sequences.

Video-Language Transformer. As transformers are more and more widely used in the visual field, some cross-modal pre-training methods based on language and vision for video tasks have been proposed,e.g., VideoBERT[6] is the first work in Video-Language Pre-training to learn a bidirectional joint distribution of visual and linguistic token sequences based on the bert model for action classification and video captioning tasks. CBT[16] divides video input and text input into two branches, and the two sequences go through the Attention module of the crossed Transformer for multi-task pre-training. ActBERT[24] enables self-supervised learning of joint representations of video and text from unlabeled data. HERO [10] encodes multimodal input in a hierarchical structure. ClipBERT [9] cancels the backbone of video feature encoding and uses an end-to-end method to train the pre-training model. The model structure of VLM[18] is an encoder with only one BERT, and the input of the model is the video feature sequence extracted by S3D[17] and the text feature sequence. VideoCLIP[19] uses temporally overlapping positive sample pairs and negative samples from the nearest neighbor retrieval for comparative learning, followed by zero-shot transfer, which surpasses previous work in many downstream tasks.

3 OUR METHOD

The pipeline of our method is shown in Figure 2. First of all, since the relationship between characters is more stable and the interaction is more changeable, we first cluster provided clips with the help of LGSS [13] to obtain a longer video, which will be called scene in the following. We predict interactions based on clips and relationships based on scenes, and then extract the features of clip and scene, including video features, text features, audio features, entity features and pose features. We concatenate all the features to represent the video. In order to reflect the influence of interaction and relationship on each other, we take the average of clip features which consist a scene as scene feature, and concatenate scene feature to each clip feature to update clip feature. In addition, in order to introduce the influence of sentiment on relationship and interaction prediction, we also use clip feature to predict sentiment as another branch of the multi-task framework. Finally, we perform a linear mapping to transfer the fusion feature into a fixed length. We concatenate the mapped features and category features as the input of our similarity calculation network to get the similarity score. The similarity score is taken as the final prediction result to help developing the knowledge graph which is used to answer the queries.

3.1 Multi-modal Feature Extraction

Visual Feature We firstly track entities with the help of tracking algorithms and face recognition approaches which is the same as our previous method[22]. Then we sample each video and reshape the bounding boxes of the tracked entities as visual input. Finally, different video feature extraction models, such as I3D [3] and VLM[18], will extract video and entity features using the inputs.

Text Feature We use DownSub[1] an automated subtitle extraction tool, to extract video subtitles, and then we extract the feature of subtitles as text feature with the help of Bert[7].

Audio Feature We first export the video into audio format and then calculate the different order mfcc features as audio features.

Pose Feature We first use OpenPose[12] to obtain human pose points, and then calculate the intersection between pose pixels and tracking bounding boxes to obtain pose tracking results. We then calculate the joint motion feature as pose feature between two entities. Specifically, joint motion feature is the euclidean distance among different pose points between adjacent frames, which is calculated as follows

$$D(i, j; t_1, t_2) = ||p^x_{i,t_1} - p^y_{j,t_2}||, \tag{1}$$

where x, y refer to different entities, t_1 and t_1 means adjacent frames, i and j are any points of poses.

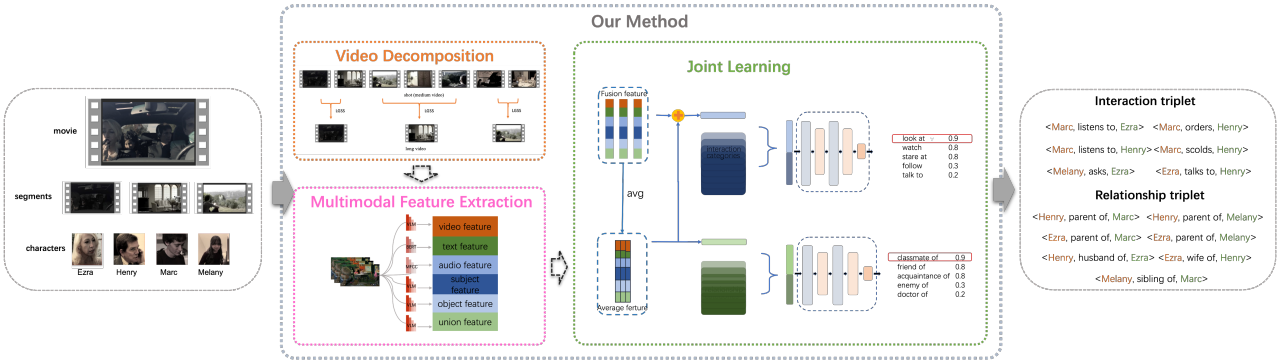

Figure 2: The pipeline line of our method. We use a joint learning framework to simultaneously perform multiple prediction tasks with multimodal features. We firstly cluster clips into longer time videos. Then we use VLM, Bert and MFCC to extract visual, text and audio features. Finally we fuse the multimodal fatures to calculate the similarity scores between videos and categories. Based on the prediction result, we can answer the queries of DVUC.

3.2 Video Language Transformer

Transformer has recently been applied in many areas and achieved great effect, which contributes to the understanding of global information with its advanced attention mechanism. VLM[18] uses the structure and training method of Bert[7], which is a classic transformer, for reference. It predicts video and text mask respectively by combining the features of video and text. By this way, it can learn the influence between video and text and better represent the relation between video and text. In view of the cross-domain characteristic of DVUC queries, we choose VLM as backbone to export visual features and text features, then input them to the prediction model for feature fusion, and finally train the classifier to obtain the results.

In addition, since the question and answer are both in the form of text, we also attempt to directly input video and choice to VLM, and then export visual features and text features and match them to get the answers.

3.3 Query Answering

Movie-level. Q1:Find all path. We set characters as points and relationships as edges, construct a graph for each movie, and use the graph search algorithm to query the given source and target. Q2:Fill in space. We traverse the edges of movie graphs, obtain candidate edges that match the queries, and sort them according to the scores generated by our method. Q3:Choice. 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. These queries are split into two groups where Q1 is group one and Q2 and Q3 comprises group two.

Scene-level. Q1:Find the unique scene. We traverse each knowledge graph, add matching interactions to the candidate list, count the number of matches, then sort the choices. Q2: Fill in space. We traverse the edges of the entity-relation graphs, match the interactions where the predicate and the object are exactly the same, count and sort the triplets. Q3/Q4: Find previous/next interaction. We divide the scene into smaller slices, each of which contains one interaction, and we give predictions by sequence. Q5: Match description. For each scene in the options, we match the descriptions one by one according to the number of occurrences of entity and object. Q6: Find sentiment.we choose the sentiment with the highest predicted score as the answer. These queries are also split into two groups where Q1, Q2 and Q3/Q4 comprises group one and Q5 and Q6 comprises group two.

Compared with last year, the main improvements we have made are that in Q3/Q4 of the scene-level, where the prior knowledge is added to assist in judging the sequence of interactions, and in Q5, the strategy of matching descriptions from the scene one by one is adopted. Besides, we also fuse the knowledge in queries to help answering movie-level questions.

As for the prediction of start time and end time, considering the difficulty of the task, we temporarily answer it with the start time and end time of the scene, and will conduct some more in-depth research in the following work.

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 [5]. HLVU dataset has 20 movies in total, and each of them last for ninety minutes on average. There are 14 movies of the HLVU dataset are development data and the remainings are test data.

In our experiments, we evaluate the performance of interaction detection using metric *Recall@k*, which is usually applied in detection tasks. The metric *Recall@k* is computed by

$$Recall@k = \frac{TP_k}{TP_k + FN_k}, \tag{2}$$

where TP_k and FN_k denote the number of correct label predicted and unpredicted in the top k confident predictions, respectively. k is set to 10 for the evaluation of interaction prediction performance.

As there are many multiple choice questions, we also use the correct answer rate as the metric. The correct answer rate is difined as follows

$$Accuracy = \frac{N_{correct}}{N_{total}}, \tag{3}$$

$N_{correct}$ is the number of correct answers and N_{total} is the total number of questions. There are four movies for validation, and each movie have four questions about sentiment and ten questions about description.

4.2 Multiple modality Feature

We experiment on different combinations of video, text, audio and pose features to verify the influence of multiple modality features. The results are shown in Table 1 where V means video feature, T means text feature, A means audio feature and P means pose feature.

Among them, the varient of the text combined with pose feature performs best, and the effect is better than that of text feature alone, indicating that the added pose feature is effective. Besides, the varient of adding video feature to text and pose feature becomes worse. We think this is because the difference between video feature and posture feature is too large, which brings noise instead. This can also be observed from the huge decrease between the effectiveness of the varient using video alone and the varient combining video and pose features. The varient with audio feature is also worse in the same way, which we think is caused by the large difference between different features.

4.3 Video Language Transformer

To conduct sentiment prediction, we verify the validity of VLM, as shown in Table 2. Our baseline takes I3D as the video feature extraction model and extracts the subtitle feature with Bert, and then input the fusion of them into a classifier to predict sentiments. We then replace the video feature extraction model of baseline into VLM as variant 1. Variant 2 uses VLM as video and text feature extraction model. In addition, we directly use VLM to extract video and choice features and then match them to answer queries as variant 3. We also match the subtitle features extracted by VLM with choice features as variant 4 which only involves text domain.

The experimental results show that variant 3 has the best performance, which demonstrates the effectiveness of using VLM directly to extract video and choice feature and then matching them. We think that it is because there exist few shot, zero shot and data bias problems in the dataset. Besides, simple classifier training is not as effective as the pre-trained model which has obtained considerable cross-domain knowledge.

4.4 Query Answer Matching Strategies

We focus on description-related tasks as video text matching tasks and experiment on different matching strategies. Firstly, we use deep learning network for feature extraction, and then calculate the similarity between video and description feature as the matching score. Variant 1 uses VLM to extract features of videos and choices, and then calculate their similarity to answer queries. Since the entities appearing in each video can be recognized through multiple object tracking, and the names of characters usually appear in the description of the video, we also match the entity appearing in the video with the description to obtain the answer. Variant 2 is to match different videos with one description to choose the correct video related to the description on the basis of the entity matching strategy. It can also be observed that in the description related questions, the ten video choices correspond to the ten description questions one by one, so we match different descriptions with one video to obtain the description that best describes the video as variant 3. The results of the experiment are shown in Table 3, and variant 3 achieves the best performance, indicating that in the query about video description, the effect of matching entities is better than that of matching features extracted by some deep learning methods.

Table 1: Experiments on multiple modality features. We calculate *Recall*@10 of interaction prediction results to evaluate the performance of different combinations of features.

	All	Bagman	Manos	RoadToBali	TheIllusionist
V	4.70	3.40	12.70	3.10	2.60
V+P	0.90	1.10	2.60	0.00	0.40
V+T	10.90	**9.80**	**28.0**	10.90	**4.40**
V+T+A	7.00	2.90	16.90	8.00	3.10
V+T+A+P	2.30	1.40	10.10	0.30	0.90
T	8.50	3.70	24.30	8.50	2.60
T+P	**11.30**	6.00	27.00	**12.40**	**4.40**

Table 2: Experiments on sentiment query answering using VLM. We calculate accuracy of the answers to evaluate the performance of different variants.

	All	Bagman	Manos	RoadToBali	TheIllusionist
V_{I3D}+T_{Bert}	0.19	0	0.25	0.25	0.25
V_{VLM}+T_{Bert}	0.19	0	0.25	0.25	0.25
V_{VLM}+T_{VLM}	0.19	0	0.25	0.25	0.25
V_{VLM}	**0.44**	**0.25**	**0.5**	**0.5**	**0.5**
T_{VLM}	0.25	**0.25**	0.25	0.25	0.25

Table 3: Experiments on matching strategies for description query answering. We calculate accuracy of the answers to evaluate the performance of different variants.

	All	Bagman	Manos	RoadToBali	TheIllusionist
VLM	0.15	0.10	0.0	0.40	0.10
$Entity_{D2V}$	0.20	0.10	0.20	0.40	0.10
$Entity_{V2D}$	**0.40**	**0.30**	**0.50**	**0.50**	**0.30**

5 CONCLUSIONS

In this paper, we introduce pose feature to the multi-modal joint learning solution to improve the performance of interaction prediction. To facilitate understanding videos, VLM which learns cross-domain knowledge is introduced. We also design different matching strategies between questions and answers. The experimental results prove that our work brings performance improvement for different tasks. And we finally rank first for group one movie-level queries and rank third for both of group one and group two scene-level queries.

ACKNOWLEDGMENTS

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

REFERENCES

[1] DownSub. https://downsub.com/.
[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In

Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 3686–3693.

[3] 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.

[4] Anoop Cherian, Julien Mairal, Karteek Alahari, and Cordelia Schmid. 2014. Mixing body-part sequences for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2353–2360.

[5] 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.

[6] 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.

[7] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.

[8] Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. 2020. Learning Interactions and Relationships between Movie Characters. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9849–9858.

[9] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7331–7341.

[10] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200* (2020).

[11] Diogo C Luvizon, David Picard, and Hedi Tabia. 2018. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5137–5146.

[12] Daniil Osokin. 2018. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. *arXiv preprint arXiv:1811.12004* (2018).

[13] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A Local-to-Global Approach to Multi-modal Movie Scene Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10146–10155.

[14] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *ACM international conference on Multimedia*. 1300–1308.

[15] Haoquan Shen, Shoou-I Yu, Yi Yang, Deyu Meng, and Alexander Hauptmann. 2014. Unsupervised video adaptation for parsing human motion. In *European Conference on Computer Vision*. Springer, 347–360.

[16] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 1906. Contrastive Bidirectional Transformer for Temporal Representation Learning. 2019a. *URL http://arxiv. org/abs* (1906).

[17] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*. 305–321.

[18] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021. VLM: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996* (2021).

[19] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084* (2021).

[20] 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.

[21] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. 2012. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 28–35.

[22] Beibei Zhang, Fan Yu, Yaqun Fang, Tongwei Ren, and Gangshan Wu. 2021. Hybrid Improvements in Multimodal Analysis for Deep Video Understanding. In *ACM Multimedia Asia*. 1–5.

[23] 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*.

[24] Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8746–8755.