DCU ADAPT at TRECVID 2022: Deep Video Understanding challenge

Anastasia Potyagalova and **Gareth J. F. Jones**
ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland
anastasia.potyagalova2@mail.dcu.ie
Gareth.Jones@dcu.ie

Abstract

We describe details of our participation in the TRECVID 2022 Deep Video Understanding challenge. For this task, we generated text descriptions for each video segment based on the entities recognised in the segment. We then built a knowledge graph, which clarifies the connections between characters, locations and other entities. This solution gave acceptable results for the scene-level track but was ineffective for the movie-level track queries.

1 Introduction

Our approach to both of the tasks in the Deep Video Understanding challenge was based on feature extraction and text caption generation for the segmented episodes. Our scene-level track solution is based on the detection of actions in the sub-clips and selecting synonyms from the list of recognized actions. The movie-level track solution uses pre-detected entities of sub-clips, activities and connections between them, building a knowledge graph for the whole film [1].

Text caption generation for the segmented episodes was carried out by building connections between visual content and corresponding text descriptions. To do this, we compared extracted features from reference images and selected frames from each segment. For this case, if an entity appears more often than an empirically predefined threshold (selected based on an experimental trade-off between correct and false recognition of entities), it is added to the detected entities list. After this, we prepare a list of the possible detected actions defined by the PyTorch X3D_M model[5]. X3D_M models provide the list of possible detected actions with their associated probability. For our solution, we took the top-15 predicted actions for each video segment. For our investigation we prepared two different lists for each video segment by using different models, X3D_M and X3D_S; both of these provide high accuracy on various datasets [5], [4]. Finally, we compare scene-level queries with each of the recognised list of activities and segments containing actions (or synonyms) present in the query are selected as answers.

The following steps were used to construct the knowledge graph for each movie. Before the construction of the graph, we have identified a list of detected entities and a list of detected actions for each video fragment. Therefore, it was possible to generate text descriptions for each video episode containing the entities recognized before. After this step, we build the knowledge graph based on generated text annotations [6]. Still, this graph did not help answer the questions about deeper relations between characters as it was necessary to answer movie-level queries.

2 Approach

The scene-level track consists of several sub-tasks: find the unique scene with a predefined list of characters' interactions and the next and previous interactions between selected characters. The first step of our solution for this track included the following steps:

- Detect all actions of the characters by using the PyTorch X3D models. Following this step, we have a list of the characters acting in the segment and a list of actions detected in these segments. These data will be used in scene-level and movie-level tracks.
- Search for coincidences or synonyms using the Gensim library methods[7].

The movie-level track also consisted of sub-tasks:

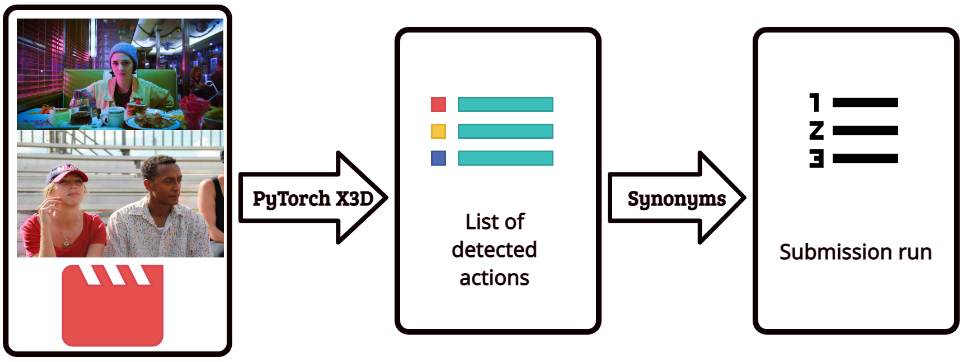

Figure 1: Method for scene-level track

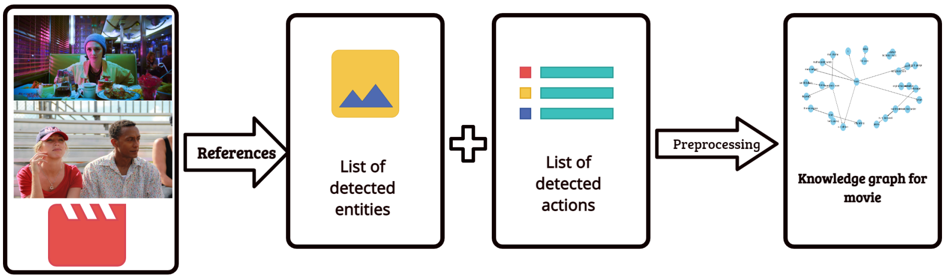

Figure 2: Method for movie-level track

- Build the knowledge graph for each movie from the dataset
- Fill the part of the graph question by filling in the blank nodes and answer questions about a person's relationship with other persons or entities; choosing one answer from multiple options as necessary.

  The movie-level solution could be separated into the following steps:
- Build the distribution for entities for each video segment by using the feature extraction algorithm provided by EfficientNetV2 [8][3]. This step was necessary in order to generate brief text descriptions for the segments for use later for the building knowledge graph.
- Generate the text description for each video segment, using the detected entities distribution and recognized actions from the scene-level solution.

After completing these steps, we have a list of entities and actions for each video segment. This means that it is possible to create text captions for the sub-episodes. Then, after this step, we have a larger piece of text describing the individual episodes, one by one. Following this, we create the knowledge graph according to the standard algorithm [2]:

- Coreference resolution
- Entity recognition
- Entity linking
- Co-occurrence graph

3 Results

Table 1: TRECVID-2022 results

Scene-level task results		
Team	Points	Percentage
Columbia_1	11.35	15.8%
Adapt	11.00	15.3%
Columbia_2	8.35	11.6%
WHU_NERCMS_1	8.00	11.1%
WHU_NERCMS_2	2.25	3.1%

The average results for the team and scene-level track are presented in Table 1. Table 2 shows results for the Adapt team for each video from the provided dataset.

Unfortunately, our approach, using brief text descriptions, is unreliable for knowledge graph creation.

Table 2: Detailed scene-level results

Scene-level results for each movie		
Movie	Points	Average
Calloused hands	1.00	0.083
Chained for life	1.00	0.083
Liberty kid	3.00	0.25
Like me	3.00	0.25
Little rock	0.0	0.0
Losing ground	3.00	0.25

4 Conclusions

Our approach to the scene-level track gave reasonably good results (15.3% accuracy or 11.00 points). However, a more accurate action detection model could be used to improve this result, and it may be useful to build a connection between detected actions and timestamps; the timeline of recognized actions will help to find the answer for the first scene-level query. Unfortunately, we found that the movie-level approach was not accurate enough to answer the defined questions, so the movie-level answers were not submitted. Improving the steps in construction of the knowledge graph processing may be helpful to the movie-level track[9].

5 Acknowledgements

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224 and partially as part of the ADAPT Centre at DCU (Grant No. 13/RC/2106_P2) (www.adaptcentre.ie). For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] G. Awad, K. Curtis, A. A. Butt, J. Fiscus, A. Godil, Y. Lee, A. Delgado, J. Zhang, E. Godard, B. Chocot, L. Diduch, J. Liu, Y. Graham, , G. Quénot, An overview on the evaluated video retrieval tasks at trecvid 2022, in: Proceedings of TRECVID 2022, NIST, USA, 2022.

[2] X. Chen, S. Jia, Y. Xiang, A review: Knowledge reasoning over knowledge graph, Expert Systems with Applications 141 (2020) 112948.
URL https://www.sciencedirect.com/science/article/pii/S0957417419306669

[3] F. Chollet, et al., Keras (2015).
URL https://github.com/fchollet/keras

[4] X. Du, Y. Li, Y. Cui, R. Qian, J. Li, I. Bello, Revisiting 3d resnets for video recognition, arXiv preprint arXiv:2109.01696.

[5] C. Feichtenhofer, X3d: Expanding architectures for efficient video recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[6] L. Mahon, E. Giunchiglia, B. Li, T. Lukasiewicz, Knowledge graph extraction from videos, CoRR abs/2007.10040.
URL https://arxiv.org/abs/2007.10040

[7] R. Rehurek, P. Sojka, Gensim–python framework for vector space modelling, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3 (2).

[8] M. Tan, Q. V. Le, Efficientnetv2: Smaller models and faster training.
URL https://arxiv.org/abs/2104.00298

[9] D. Vasile, T. Lukasiewicz, Learning structured video descriptions: Automated video knowledge extraction for video understanding tasks, in: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", Springer, 2018, pp. 315–332.