Semantic Understanding and Evolving Interaction Tracking in Long-form Multimodal Datasets

Note: This is an early draft, the camera-ready version will be available in March 2023.

Vishal Anand [1,3], **Yifei Dong** [1,2], **Raksha Ramesh** [1,2],
Zifan Chen [1,2], **Yun Chen** [2], **Linquan Li** [1,2], **Ching-Yung Lin** [1,2]
[1]Columbia University, [2]Graphen Inc.
[3]Microsoft Corporation
{vishal.anand, yd2616, rn2486, zc2628, ll3466, c.lin}@columbia.edu
{yunchen, cylin}@graphen.ai

Abstract

We share our analysis of how language interactions evolve in long-form multimodal datasets. This draft describes the experiment setups along with learning and observations that resulted in leading the scoreboard at the NIST DVU challenge in 2022. The draft is currently being revised for the camera-ready version to be made available on March 2023.

1 Introduction

After our recent work on MultiModal Language Modelling (Anand et al., 2021; Ramesh et al., 2022), we focused on User-interaction Mapping (Section 2.1), Segmenting User Stories (Section 2.2), Auto-Analysis System(Section 2.3), Prompt Variation (Section 2.4), Large Language Model Analysis (Section 2.5) and Model Correctness Analysis (Section 2.6).

2 Methods

2.1 User-interaction Mapping

In our early work, we developed a novel approach for multi-entity tracking, which shows promising results in mapping entities' names with their figures on the frame - especially when they are facing away from the camera or when less than half of their faces are visible. Also, face embeddings are computed from five facial land-mark points: eye-left, eye-right, nose, mouth-left, & mouth-right using additive angular margin-loss to perform face recognition. After comparing the results from implementing multi-entity tracking and only face recognition, respectively, we have some interesting findings. First, multi-entity tracking provides valid predictions on several frames that face recognition fails to predict. Also, face recognition works on some frames that multi-body tracking fails to predict. For some frames containing both predictions from multi-entity tracking and face recognition, the results can be quite different. So we set up experiments to figure out which method performs better in this task and how to find a better method by the cross reference of multi-entity tracking and face recognition. In general, there will be four sets of experiments:

- Only use multi-entity tracking
- Merge multi-entity tracking and face recognition with the same priority
- Merge multi-entity tracking and face recognition with multi-entity tracking prioritized
- Merge multi-entity tracking and face recognition with face recognition prioritized

The merging process here is based on one principle if one frame only has one set of valid predictions from either multi-entity tracking or face recognition, we remain this set of predictions as the final prediction for this scene. Based on this standard, case 2 here means that if frame A from movie B has prediction "entity_a" from multi-entity tracking and prediction "entity_b" from face recognition, we keep both predictions "entity_a" and "entity_b" for frame A. For case 3, in the same situation, we just save the prediction "entity_a" for frame A. For the last case, we just save the prediction "entity_b" for frame A.

From the results of these four sets of experiments shown in Tables 1, 2, 3, 4, 5, 6, 7, 8, we can find that Case 1: only use multi-entity tracking gets the best results on M1-MRR with 60.9% upper-limit and 29.4% lower-limit. Case 2: Merge multi-entity tracking and face recognition with the same priority gets the best results on S4-Acc with 22.8%. Case 3: Merge multi-entity tracking and face recognition with multi-entity tracking prioritized gets the best results on S1-MRR with 16.0% and on M2-Acc with 21.1%. Case 4: Merge multi-entity tracking and face recognition with face recognition prioritized gets the best results on S3-Acc with 24.2%.

Movie	S1-MRR	S3-Acc	S4-Acc
The_Big_Something	5.8	9.1	9.1
honey	21.2	21.7	22.5
shooters	9.8	26.3	42.1
Huckleberry_Finn	8.2	19.7	21.1
sophie	16.6	27.8	33.0
time_expired	5.4	18.5	16.3
spiritual_contact	11.1	20.8	20.8
Valkaama	28.1	30.8	30.8
Nuclear_Family	28.4	28.6	14.3
SuperHero	20.8	11.1	0.0
Average	15.5	21.4	21.0

Table 1: Training Evaluation (percentage) on scene level tasks with only body tracking

Movie	M1-MRR-U	M1-MRR-L	M2-A
Manos	46.6	20.2	25.5
Road_to_bali	45.0	18.9	19.3
Bagman	39.6	15.2	16.8
honey	73.1	29.3	32.8
shooters	56.0	34.0	12.1
Huckleberry_Finn	71.3	25.9	15.0
sophie	57.3	24.5	23.5
spiritual_contact	51.3	29.3	13.0
Valkaama	51.2	37.0	28.9
Nuclear_Family	100	52.1	9.5
SuperHero	78.6	37.0	28.9
Average	60.9	29.4	20.0

Table 2: Training Evaluation (percentage) on movie level tasks with only body tracking

Movie	S1-MRR	S3-Acc	S4-Acc
The_Big_Something	12.4	9.1	9.1
Honey	18.5	27.5	26.7
Shooters	11.6	29.0	36.8
Huckleberry_Finn	6.6	19.7	23.9
Sophie	11.7	21.7	21.7
Time_Expired	5.7	18.5	15.2
Spiritual_Contact	12.9	41.5	35.9
Valkaama	19.9	53.9	38.5
Nuclear_Family	14.9	7.1	7.1
Superhero	25.3	0.0	0.0
Average	14.0	22.8	21.5

Table 3: Training Evaluation (percentage) on scene level tasks with face recognition and head tracking equally merged

Movie	M1-MRR-U	M1-MRR-L	M2-A
Manos	44.9	20.3	26.5
Road_to_bali	41.5	18.9	26.0
Bagman	43.6	15.2	22.1
Honey	54.8	29.3	25.9
Shooters	55.6	34.0	6.1
Huckleberry_Finn	65.3	25.9	13.3
Sophie	46.0	24.5	24.7
Spiritual_Contact	55.4	29.3	21.7
Valkaama	51.2	37.0	20.0
Nuclear_Family	100	52.1	9.5
Superhero	79.8	37.0	23.7
Average	58.6	29.4	20.1

Table 4: Training Evaluation (percentage) on movie level tasks with face recognition and head tracking equally merged

Movie	S1-MRR	S3-Acc	S4-Acc
The_Big_Something	5.8	9.1	9.1
Honey	21.5	24.2	22.5
Shooters	10.0	26.3	36.8
Huckleberry_Finn	8.2	19.7	21.1
Sophie	15.5	17.5	21.7
Time_Expired	5.4	18.5	16.3
Spiritual_Contact	13.8	37.7	35.9
Valkaama	30.5	38.5	23.1
Nuclear_Family	28.4	28.6	14.3
Superhero	20.8	11.1	11.1
Average	16.0	23.1	21.2

Table 5: Training Evaluation (percentage) on scene level tasks with body tracking prioritized

Movie	M1-MRR-U	M1-MRR-L	M2-A
Manos	49.3	20.2	23.5
Road_to_bali	37.8	18.9	18.0
Bagman	23.1	15.6	19.8
Honey	64.8	29.3	29.3
Shooters	58.3	34.0	12.1
Huckleberry_Finn	42.4	25.9	13.3
Sophie	43.6	24.5	22.4
Spiritual_Contact	40.7	29.3	19.6
Valkaama	51.0	37.0	30.0
Nuclear_Family	100	52.1	9.5
Superhero	78.6	37.0	34.2
Average	53.6	29.4	21.1

Table 6: Training Evaluation (percentage) on movie level tasks with body tracking prioritized

Movie	S1-MRR	S3-Acc	S4-Acc
The_Big_Something	10.6	45.5	27.3
Honey	13.9	25.0	21.7
Shooters	13.9	31.6	36.8
Huckleberry_Finn	5.6	18.3	22.5
Sophie	12.8	19.6	23.7
Time_Expired	3.7	19.6	18.5
Spiritual_Contact	13.7	26.4	18.9
Valkaama	21.1	30.8	15.4
Nuclear_Family	21.4	14.3	14.3
Superhero	10.0	11.1	11.1
Average	12.7	24.2	21.0

Table 7: Training Evaluation (percentage) on scene level tasks with face recognition prioritized

Movie	M1-MRR-U	M1-MRR-L	M2-A
Manos	45.2	20.2	21.4
Road_to_bali	48.5	18.9	23.3
Bagman	28.1	15.8	15.3
Honey	64.8	29.3	27.6
Shooters	58.3	34.0	15.2
Huckleberry_Finn	50.6	25.9	16.7
Sophie	58.3	34.0	15.2
Spiritual_Contact	41.0	29.3	17.4
Valkaama	51.7	37.0	20.0
Nuclear_Family	100	52.1	9.5
Superhero	65.5	37.0	7.9
Average	55.6	29.5	17.6

Table 8: Training Evaluation (percentage) on movie level tasks with face recognition prioritized

2.2 Segmenting User Stories

In our early work, we used i-frame method for key frame extraction, which is to grab the 5th frame in every ten frames. The disadvantage of this method is obvious a great number of frames that only contain location information are captured. As the goal of Deep Video Understanding is to study the interaction and relationship between human_entities, these frames are unrelated and of little value to us.

So we develop a multi-entity-frame method for key frame extraction, which is to implement face recognition on movie scenes first and select the frames containing two or more entity_faces. Between two capture, this process will sleep for 21 frames. In real practice, we find that in the case shown in 1 and 2, simply selecting a multi-entity frame cannot guarantee that we have crawled all valid information. 1 is the 3670th frame in Bagman Scene5, while 2 is the 3672th. In this particular case, person_A and person_B in 1 is having intense discussions with person_C in 2. As they are sitting in different corners of the room, each time person_C appears alone in the shot. Thus, only selecting the frame with two or more entities can only capture the frames that contain person_A and person_B and fail to capture useful information for person_C in this case, which can lead to a wrong prediction result that person_A is having intense discussions with person_B. To avoid such cases, we improve our multi-frame method. For each keyframe captured, we merge the face recognition results of a range of frames(+/- 5) with the original face recognition results of that keyframe and use this merged result as the final face recognition result for it. The change in the key frame extraction method leads to the change of method for Entity mapping in clip prediction.

In our early work, the method for Entity mapping in clip prediction is shot-based: After obtaining the predictions for the semantic attributes, we use the multi-body tracking framework to obtain the source and target entity between whom the interaction occurs in every keyframe. Furthermore, for some shots with missing entities, we use a range of shots(+/- 1) to infer the entity within the shot, i.e., if an entity occurs in the previous and next shot they are most likely to occur in the current shot too. The reason why we use the above shot-based method is that we cannot guarantee that each keyframe contains human_entity information for entity mapping. In the current work, as we develop a multi-entity-frame method for key frame extraction, the method for Entity mapping in clip prediction is frame-based. For example, all prediction results in frame A are mapped with all the human_entities that appear in frame A.

Figure 1: Bagman-5-3670

Figure 2: Bagman-5-3672

2.3 Auto-Analysis System

Proper data visualization helps us to understand the results obtained from the testing process more easily. Likewise, it is important to integrate and functionalize these outputs on a public platform so that more people can understand the results of the entire experiment. This study will discuss the integration of the data obtained after the test and the design of the corresponding one-stop-shop website (OSS). In the design of OSS, we will discuss that the back-end platform generates data and integrates the design while generating the website and displays the results on the front-end platform in the form of data visualization. In the data integration section, we will discuss real-time data generation and aggregation into a single folder for visual inheritance. In the visualization part, it will contain the scene-level knowledge graph and the question results of each movie.

In the knowledge graph presentation of the scene, the main sources of information are the following:

- Face tracking of characters
- Nodes: object name, node's type, character's face tracking

- Edges: relation type, relation, source, target
- Fine-grained inference of interactions, sentiments, location, and emotions through image-language model

In order to enhance the expression of information, we use the above components for visualization. Furthermore, to make the results easier to operate, we developed a user interface for researchers to understand the model's results. As shown in Figure 3, by creating a two-level drop-down menu, users can choose the movies and scenes they want to visit. When the user selects the option, the corresponding scene video and relationship diagram will be displayed. Through the processing of data visualization, we can intuitively understand the interaction between people, locations, and current emotions through the results.

On the other hand, the evaluation for the query and answering visualizations are also included on the home page of the website. We convert the evaluation result into a histogram, the x-axis is the movie title and question, and the y-axis is the accuracy. The user can access the specified results and compare them by selecting the items. In Figure 4 (a), when we click on the evaluation page, we will see six queries and answering results in one histogram. Then, as shown in Figures 4 (b) and (c), by choosing the option with the selecting box, we can leave the items we need and remove the unneeded items.

2.4 Prompt Variation

To have a better performance on scene description. Locations are included in each prompt as the input for the text encoder. This idea comes from attaching locations in every prompt as common sense for a higher confidence score. This method is applied on person to person relations and person-to-location relations under the following format person-location relationship: A photo of a person in {location} who {relationship} at person-person relationship: A photo of a person in {location} who is {relationship} Sentences without the prompt are in the original format. The locations and object entities are localized within scenes using SIFT based on feature matching to handle various scales and crops. Note that when dealing with batch inputs, keyframes have multiple locations. To avoid the overwhelming size of prediction labels, we adopted prediction without batch processing.

Movie	M1-MRR-U	M1-MRR-L	M2-A
Manos	49.3	20.2	23.5
Road_to_bali	37.8	18.9	18.0
Bagman	23.1	15.6	19.8
Honey	64.8	29.3	29.3
Shooters	58.3	34.0	12.1
Huckleberry_Finn	42.4	25.9	13.3
Sophie	43.6	24.5	22.4
Spiritual_Contact	40.7	29.3	19.6
Valkaama	51.0	37.0	30.0
Nuclear_Family	100	52.1	9.5
Superhero	78.6	37.0	34.2
Average	53.6	29.4	**21.1**

Table 9: Training Evaluation (percentage) on movie level tasks without location prompt

Movie	M1-MRR-U	M1-MRR-L	M2-A
Manos	44.2	20.2	20.4
Road_to_bali	35.1	18.9	17.3
Bagman	26.0	15.6	13.0
Honey	69.0	29.2	29.3
Shooters	60.8	34.3	12.1
Huckleberry_Finn	56.9	25.9	8.3
Sophie	36.5	24.5	20.0
Spiritual_Contact	41.0	29.3	21.7
Valkaama	52.9	37.0	33.3
Nuclear_Family	100	52.1	14.3
Superhero	78.6	37.0	23.7
Average	**54.6**	29.4	19.4

Table 10: Training Evaluation (percentage) on movie level tasks with location prompt

As shown in 10 and 9, this experiment with extra locations in prompts show a higher score on M1-MRR-upper metric while lower performance on M2-Acc. From a high-level view, the overall performance is not much different from the original prompt. Based on the fact that both image and description are fed into the CLIP model. The extra location in the text can also be extracted by the image encoder. Including locations in prompts may not lead to significant progress.

2.5 Large Language Model Analysis

Apart from scenes and motions, information is derived from the dialogues. Knowledge gained directly from conversations, in a sense, is more elaborated than the inference of scenes and frames (imagine a person reading the dialog subtitles versus another person watching a muted and unsubtitled movie). From this notion, we feed the subtitles to a text model for relationship extraction.

We refer to the approaches in Dialogue-Based Relation Extraction. Given a dialogue $D = s_1 : t_1, s_2 : t_2, \ldots, s_m : t_m$ and argument pair (a_1, a_2), where s_i and t_i denote the speaker ID and text of the i^{th} turn, respectively, and m is the total num-

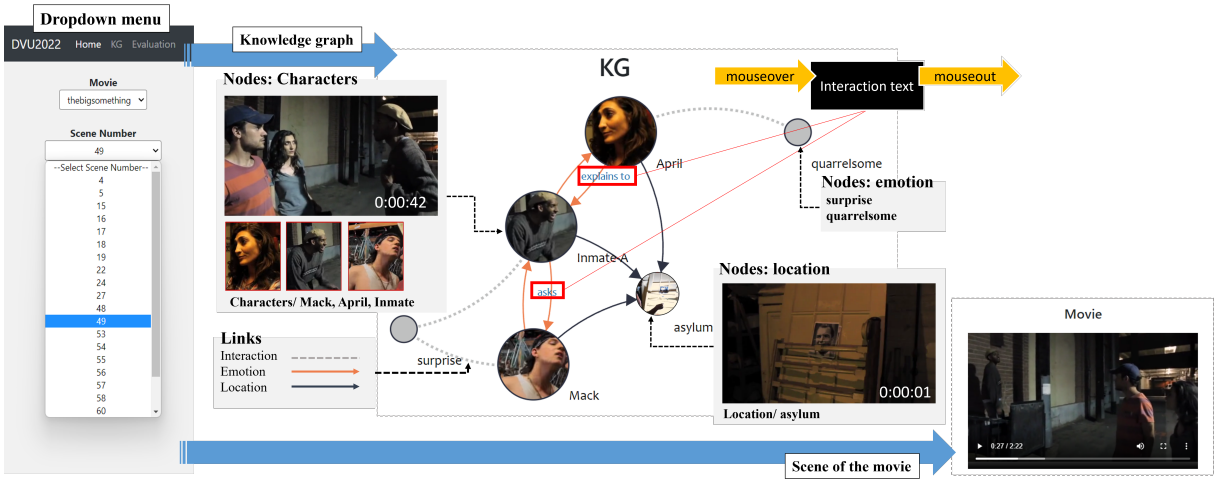

Figure 3: Use the drop-down menu to obtain the Knowledge Graph and scene video for The Big Something for scene 49

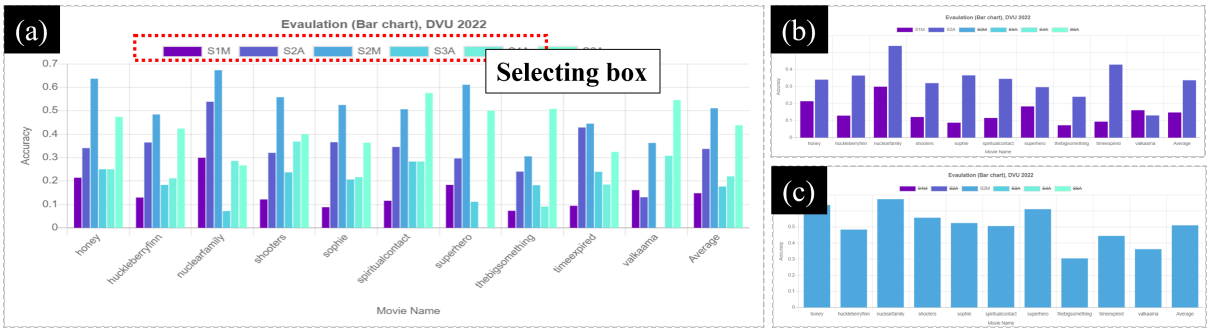

Figure 4: (a) The histogram of the accuracy of each movie under each query, (b) Choose Answer1 and Answer 2 result, and (c) Only Answer 3 result

ber of turns, we evaluate the performance of approaches in extracting relations between a_1 and a_2 that appear in D.

2.6 Model Correctness Analysis

2.6.1 Scene Level Question

For scene-level questions, the first question asks us to find a specific scene based on an instruction set, and the result (average MRR score for each movie) shows a pattern with polarization. For movies *Calloused Hands*, *Chained For Life*, *Liberty Kid*, and *Losing Ground*, they have relatively low average MRR scores for this question: 0.072, 0.016, 0.046, and 0.125. However, for movies *Like Me* and *Little Rock*, they have relatively high average MRR scores for this question: 0.539 and 0.289. For the other two scene-level questions that ask the previous/next interaction after one specific interaction in a scene, we could not get a meaningful pattern from the result. The reason for this is our model does not involve any temporal factor in both input and output, and we basically "guess" the answer.

2.6.2 Movie Level Question

For the first question, we did not consider the answer could be *location*. We only used a person as our possible answer. This problem could be improved by choosing the correct answer set (person or location) before answering the question according to the type of the relation(person-person or person-location) and subject type (person or location) in the question prompt. For the second question, sometimes the question asks what the relation is between 2 people. We may give an answer that is person-location relation instead of person-person relation because the knowledge does not contain one relation between these 2 people that is in the question's choices. Then we will choose the closest relation in choice according to the relation similarity matrix if these 2 people do have a relation or give a random choice otherwise. We could improve this by filtering the choice first to omit the incorrect type of relation(i.e., person-person or person-location) according to the prompt of the question.

References

Vishal Anand, Raksha Ramesh, Boshen Jin, Ziyin Wang, Xiaoxiao Lei, and Ching-Yung Lin. 2021. Multimodal language modelling on knowledge graphs for deep video understanding. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4868–4872, New York, NY, USA. Association for Computing Machinery.

Raksha Ramesh, Vishal Anand, Zifan Chen, Yifei Dong, Yun Chen, and Ching-Yung Lin. 2022. Leveraging text representation and face-head tracking for long-form multimodal semantic relation understanding. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 7215–7219, New York, NY, USA. Association for Computing Machinery.