MultiModal Language Modelling on Knowledge Graphs for Deep Video Understanding

Vishal Anand [1], Raksha Ramesh [1,2], Boshen Jin [1,2], Ziyin Wang [1], Xiaoxiao Lei [2], Ching-Yung Lin [1,2]
[1] {va2361, rn2486, bj2437, zw2605, c.lin}@columbia.edu, [2] {xiaoxiao, cylin}@graphen.ai
[1] Columbia University, New York, NY, USA
[2] Graphen, Inc., New York, NY, USA

ABSTRACT

The natural language processing community has had a major interest in auto-regressive [4, 13] and span-prediction based language models [7] recently, while knowledge graphs are often referenced for common-sense based reasoning and fact-checking models. In this paper, we present an equivalence representation of span-prediction based language models and knowledge-graphs to better leverage recent developments of language modelling for multi-modal problem statements. Our method performed well, especially with sentiment understanding for multi-modal inputs, and discovered potential bias in naturally occurring videos when compared with movie-data interaction-understanding. We also release a dataset of an auto-generated questionnaire with ground-truths consisting of labels spanning across 120 relationships, 99 sentiments, and 116 interactions, among other labels for finer-grained analysis of model comparisons in the community.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Scene understanding**; **Knowledge representation and reasoning**.

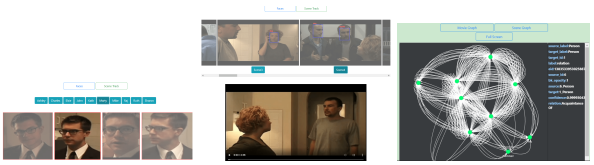
KEYWORDS

language model, transformers, slot filling, intent detection, knowledge graphs, scene description, speaker diarization

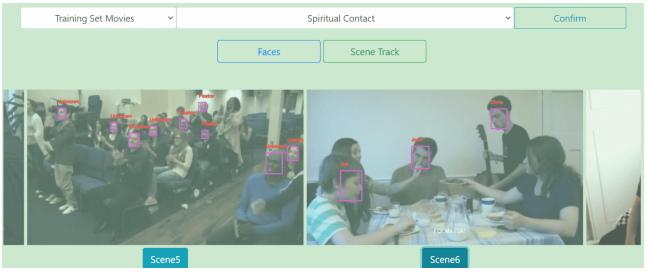
ACM Reference Format:
Vishal Anand, Raksha Ramesh, Boshen Jin, Ziyin Wang, Xiaoxiao Lei and Ching-Yung Lin. 2021. MultiModal Language Modelling on Knowledge Graphs for Deep Video Understanding. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3474085.3479220

1 INTRODUCTION AND BACKGROUND

As the natural language processing community is making increased inroads into understanding human conversations and human-like chat-bots, the research attention increased the spans across different modalities to improve understanding of natural conversations. The



(a) L-R: Face-scenes, entity-tracking-scene, knowledge-graph



(b) Entity interaction across spatial-temporal space

Figure 1: Sections of multi-domain data generated

Deep Video Understanding (DVU) dataset [5] has been explored recently by the community [1, 3, 6, 12, 14] and is highly relevant to carry out our experiments on deep-video-understanding tasks across free-form information modalities. While models attempt to merge information across modalities, one commonly faces the problem of knowledge representation in different embedding spaces, namely a label-type information would be in a different embedding space as opposed to say an image representation in a neural network. A way to deal with the problem is to tune network's hyperparameters for individual modalities. We worked with the representations identified in the paper Vishal et al. [1] and realized the transformer-based approach for decoding the network parameters could be very useful especially if a native transformer model, such as HERO [9] is trained with ample video dataset sizes. Exploration of other auto-regressive models and span-predicting models could mean performing few-shot or zero-shot learning. However, given the movie datasets are few and rare to come by, we decided to look into zero-shot transfer learning paradigms built on the basis of these two papers [1, 9], including the data-preparation and building blocks as a starting point.

2 SYSTEM

2.1 Human-Interaction Interface

To analyze multi-modal models, we created an interface currently internally hosted at Graphen Inc (fig. 1) to create, view and interact

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3479220

with knowledge bases. A link to a version with static analysis of task dataset will be released for public access here. [1]

Researchers can either upload a movie, or choose a *training movie* from the DVU-training-set. The scene-splits are then generated from the original video. For each scene's video, bounding boxes on detected faces are tracked along the video and a knowledge graph visualization is generated below the video while being processing through the language models. One can then interact with the graph by zooming, dragging or centering on nodes, and clicking specific nodes to display scene level information from the knowledge graph. Questions on relations, sentiment, interactions and question-answering are generated on the fly.

Further versions of the platform will allow researchers for sophisticated interactions and visualizations to help understand videos better, including crowd-sourcing annotation and human-suggested corrections on proposed answers and scene-boundaries and knowledge-graphs to allow for label visualization, leading to larger gold-datasets.

Researchers can dissect a given video to multiple modes of representations, such as scene-boundary generation, generate audio and transcripts, detect faces and track with bounding boxes on the faces, display names, emotions, interactions, objects, location, background as labels, and generate knowledge graphs on both movie and scene levels.

2.2 Graph Query Language - Language Modelling

We identified social relationships and interactions among entities, emotion of these entities, and sentiments in movies at scene level. With these knowledge extracted from movies, we formalized knowledge graphs that could effectively abstract movies and could be efficiently queried by *openCypher*[2] for key information (e.g. inferring relationship between two characters).

2.2.1 *Knowledge Graph Structure.*

Figure 2 describes our knowledge graph representation that has two node-types: *Person*, *Scene*, and three edges-types: *Emotion*, *Relation*, and *Interactions*. *Person* nodes' properties consists of character-names and reference IDs. *Scene* nodes have an additional property called *sentiment*, which represents a highly abstract conclusion of each scene. For edges, *Emotion* edges bridge *Person* and *Scene* nodes, indicating characters' emotions for each scene. *Relation* edges represent social relationships between any two entities in movies, and *Interaction* edges list all interactions between two people in different orders within different scenes. Since our model might render multiple interactions between two entities with different confidence-scores based on each multimodal data-segment, we can have multiple edges between them with the same order numbers and scene numbers.

[1] https://vishalanand.github.io/deep-language-multimodal-graphs/

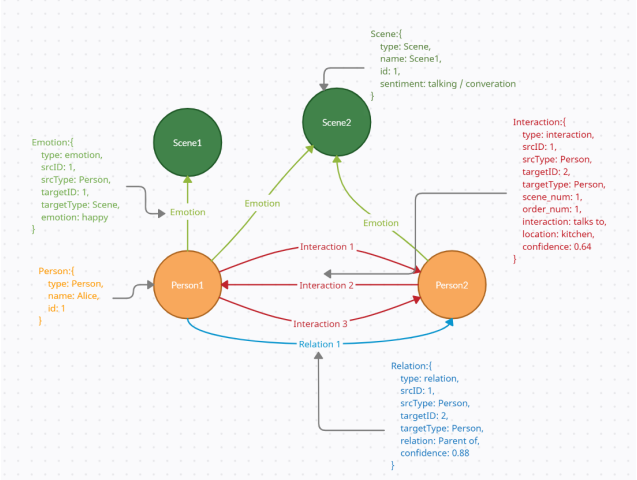


Figure 2: Knowledge Graph Structure

3 TASKS AND LANGUAGE MODEL QUERIES

Our graph query can handle various kinds of questions and these questions could be categorized into two types: movie level questions and scene level questions. For each kind of questions, our graph query has one corresponding command to fetch related information and further processes could be initiated based on returned results.

3.1 Global Scale: Movie Level Questions

(1) M1: Find relational paths between two characters
Sample: List all possible paths between *Alice* and *Bob*.
Query:

```
graphAnalytic(KG_name, "path",
{
    "depth": 6,
 "vertex_source":{"id": "1", "label": "Person"},
 "vertex_target":{"id": "2", "label":"Person"},
    "edge_labels":['relation']
})
```

(2) M2: Fill in the graph space
Sample: Which Person has Relation: Spouse Of Person:Marge?
Query:

```
queryOpenCypher(KG_name,
"match (s:Person)-[r:relation {relation:'Spouse of'}]
     -> (t:Person {name: 'Marge'})
 return s, r, t")
```

(3) M3: Question answering
Sample: How is Ms. Krabappel related/connected to Springfield Elementary?
Query:

```
queryOpenCypher(KG_name,
"match (s:Person {name: 'Ms. Krabappel'})-[r:relation]
  -> (t:Person {name: 'Springfield Elementary'})
 return s, r, t")
```

3.2 Local Scale: Scene Level Questions

(1) S1: Find Scenes based on interactions

Sample: Which Unique Scene contains the following Interactions: *explains to*, *asks*, *lies to*, *walk with*, *asks*, *talks to*, *talks to*, *talks to*

Query:

```
queryOpenCypher(KG_name,
    "match (p1: Person)
    -[i: interaction {scene_num:'%s'}]
    ->(p2: Person)
    return p1, i, p2" % (all scene numbers))
```

This iterates through all the scenes and their interactions to find the best match with prompt

(2) S2: Find Person based on scene number and interactions

Sample: Which Person in scene 13 has the following Interactions: talks to Target-Person: Princess-Lala, Source-Person: Princess-Lala greets?

Query:

```
queryOpenCypher(KG_name,
    "match (p1: Person)
        -[i: interaction {scene_num:'13'}]
        ->(p2: Person)
    return p1, i, p2"
)
```

Here, the schema iterates through all interaction pairs for the prompt to locate the *Person*

(3) S3: Find next or previous interaction between two people
Sample: n Scene 13, Prince Ken Arok watches Harold. What is the immediate next / following interaction between Prince Ken Arok and Harold, in scene 13? Query:

```
queryOpenCypher(KG_name,
  "match (p1: Person {name: 'Prince Ken Arok'})
        -[i: interaction \ {scene_num:'13'}]
        -> (p2: Person {name: 'Harold'})
    return p1, i, p2")
```

Here, the schema finds the ordering occurrence of "watch" in the language model's results and traverse to find the next interaction between these two people.

(4) S4: Find sentiment label based on scene number
Sample: In Scene 13, What is the correct sentiment label?
Query:

```
queryOpenCypher(KG_name,
   "MATCH (s:Scene \{name: '13'\}) return s")
```

(5) S5: Find scene matching with given natural-language descriptions.
(6) S6: Classify scene sentiment from a given scene.

4 METHOD AND BUILDING BLOCKS

We have a zero-shot transfer-learning model that infers and extracts information from free-form multimodal sources - text, sound, video, shot-splits, speaker-diarization, and face-tracking to create a knowledge-graph using language modelling questionnaire through slot-filling. After analysing different models based around [9, 11], we trained the model based on HERO [9] for 10000 epochs as our basis for zero-shot transfer learning owing to a lack of ample movie-data with accompanying auxiliary datasets.

HERO uses cross-modal Transformer and captures global video context through a temporal transformer, that is suited to capture multi-character interactions. As HERO is trained on HowTo100M [10] and TV datasets[8], we utilize the benchmark to investigate if the embeddings are transferrable on movie-datasets with more complex plots and evolving social dynamics. We attempt to extract the relationships and interactions between entities in a scene through an intuitive video-question answering framework. The questions we encode are as follows: "What is entity1 doing in this video?", "What is entity1's relation with entity2?" and provide the different interaction and relation class categories as prompts respectively. We localize entities through aligned speaker-diarized text, face recognition and tracking and pose these questions when the entities of interest are co-located in a scene.

Apart from the slot-filling based QA framework, we implement text-video retrieval similar to the video-subtitle matching in HERO [7] to match the scene descriptions with the scenes. We summarize the results on the scene-level queries in the following section.

5 EVALUATION AND DISCUSSIONS

Movie	Localized Metrics					
	S1	S2	S3	S4	S5	S6
Spiritual Contact	2.85	71.42	16.98	7.50	28.94	69.69
Honey	1.45	68.09	10.81	14.29	21.73	42.10
Nuclear Family	2.08	50.0	7.14	16.49	31.25	46.67
Sophie	0.06	50.53	16.49	18.87	40.00	40.91
Superhero	1.25	77.78	0.0	11.11	30.00	50.0
Huckleberry Finn	0.11	48.31	7.04	7.04	25.42	42.37
Shooters	0.34	52.0	7.89	10.52	35.00	55.0
The Big Something	0.03	66.0	18.18	18.18	31.81	36.92
Time Expired	0.02	54.29	2.17	5.43	32.00	56.76
Valkaama	0.08	73.91	30.77	23.07	37.70	50.0
Average	0.83	61.23	11.75	13.25	31.00	49.04

Table 1: Extrinsic Evaluation(%) on Tasks in Section 3.2

The extrinsic and intrinsic evaluations are listed in Table 1 and Table 2 respectively.

5.1 Automatic Data Evaluation Generator

The questions are generated on the fly on all possible sources and targets of the *unseen* knowledge-graph. *Given each source-target*

Movie	Sentiments					Interactions					
						Original Score			Normalized Score		
	MRR%	R@10	R@20	R@50	Acc	MRR%	R@20	R@50	MRR%	R@20	R@50
Spiritual Contact	48.90	55.80	60.50	81.40	46.51	5.20	7.56	37.90	8.50	34.20	50.60
Honey	28.20	52.00	68.00	84.00	16.00	4.20	17.10	42.70	4.60	23.90	41.90
Nuclear Family	9.30	23.50	29.40	58.80	0.00	1.60	4.20	20.80	6.30	33.30	66.70
Sophie	29.00	37.00	41.30	69.60	23.91	5.00	12.80	23.90	8.50	32.50	57.30
Super Hero	28.60	37.50	37.50	62.50	25.00	1.50	0.00	29.60	1.60	40.70	55.60
Huckleberry Finn	25.60	28.80	45.80	72.90	20.33	3.00	11.00	26.80	3.70	15.90	32.90
Shooters	30.90	35.00	50.00	60.00	25.00	1.50	4.40	12.40	9.40	57.50	76.10
The Big Something	29.70	33.80	40.00	64.60	26.15	1.90	6.90	15.50	8.00	36.20	63.80
Time Expired	43.10	51.40	63.50	86.50	36.48	1.50	3.60	12.10	4.50	22.10	30.40
Valkaama	38.30	50.00	52.20	63.00	32.60	1.50	1.90	11.30	12.10	60.40	77.40
Average	31.16	40.48	48.82	70.33	25.20	2.69	6.95	23.30	6.72	35.67	55.27

Table 2: Intrinsic Evaluation(%) with $\binom{n}{2}$ auto generated questions on our Model

pair, we attempt to walk our model generated output via asking language questions in a slot-filling fashion. The knowledge graph equivalence is established via OpenCypher paradigm to allow our approach to be model agnostic. This allowed us to use human-proposed knowledge-graphs to analyze how different aspects of language tasks performed for a specific movie where additional meta data can be added-on to identify weak and strong points of our multi-modal model.

5.2 Extrinsic Evaluation

The *extrinsic evaluation* is based on the actual tasks defined in Section 3.2 on the auto-generated questions. The movies analyzed in Table 1 are the training movies, which is how we have their ground truth available for evaluation.

5.3 Intrinsic Evaluation

The *intrinsic evaluation* (Table 2) are based on the validity of model-generated answers to our slot-filling questions that are leveraged to fill-in the *link-types* between the sources and targets in a given question.

5.4 Discussion and Analysis

While the extrinsic evaluation (Table 1) on given tasks gave us interesting insights into what areas our knowledge-graph did not capture well enough (S1, S3) and some areas that our model captured really well (S5, S6) in comparison, this led us to investigate why that happens to be the case.

So, we created automatic generation of $\binom{n}{2}$ intrinsic questions for the each unit of the dataset. The zero-shot transfer model fails to grasp who is speaking and to whom. It's not good enough to add face-tracking information along with other modalities here, owing to the fact that a large expanse of videos sampled for initial training does not take each person as a separate unit, but rather set of objects as a single class-type. This allows us to understand what is action is taking place in a multimodal input, but unless explicitly called out in the text or speech, our language model fails to establish interactions across the different modes. One way to solve this problem is to fine-tune our model using face-tracker based attention-heads in the decoder of transformer when the model evaluates answers to our auto-generated questions of person properties and object properties before assigning confidence scores.

Using the same analysis, we realize our model's sentiment accuracy are high because our model establishes an equivalence to language model that works very well with text-segments, thus using the text of each character's conversations to the best use, and other modes of inputs only add to it, thus leading to no unwanted negative transfer.

Moreover, we can also attribute the relatively poor performance in S3 to a distribution shift from training data, common in zero shot transfer models. Since we adapt HERO [7] trained on TVQA dataset, we find that our model captures interactions that are localized in shorter segments within a scene. But these inferred interactions are considered as false positives as the ground truth annotations are available on a high-level and not as fine-grained.

Our model performs well for (S5) matching the scenes with the scene descriptions since the model captures visual concepts, actions and descriptors well to match with the correct scene.

6 CONCLUSION

We demonstrated a multi-modal framework leveraging knowledge graphs and language model equivalence structure that infers sentiment very well, and we discover the problems on individual-entity based grounding in multi-modal frameworks, that could be considered a stretch from class-based-entity grounding. This is of specific importance for multi-modal language modelling since slot-filling mechanisms can be sensitive owing to lack of enough clearly demarcated individual-entity data across modalities, since Spoken Language Understanding (SLU) or text-only data contains ample information for individual-entity understanding. We also release the dataset on movie and scene level intrinsic evaluation to understand label specific model affinity of knowledge extraction and retention across different information modalities to help identify and consolidate multiple models' strengths into a larger ensemble that could lead to potential breakthroughs in this space.

REFERENCES

[1] Vishal Anand, Raksha Ramesh, Ziyin Wang, Yijing Feng, Jiana Feng, Wenfeng Lyu, Tianle Zhu, Serena Yuan, and Ching-Yung Lin. 2020. Story Semantic Relationships from Multimodal Cognitions. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 4650–4654. https://doi.org/10.1145/3394171.3416305

[2] Renzo Angles, Marcelo Arenas, Pablo Barceló, Aidan Hogan, Juan Reutter, and Domagoj Vrgoč. 2017. Foundations of Modern Query Languages for Graph Databases. *ACM Comput. Surv.* 50, 5, Article 68 (Sept. 2017), 40 pages. https://doi.org/10.1145/3104031

[3] Matthias Baumgartner, Luca Rossetto, and Abraham Bernstein. 2020. Towards Using Semantic-Web Technologies for Multi-Modal Knowledge Graph Construction. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 4645–4649. https://doi.org/10.1145/3394171.3416292

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., Virtual, 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[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 *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 355–361.

[6] Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. International Workshop on Deep Video Understanding. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 871–873. https://doi.org/10.1145/3382507.3419746

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*.

[9] 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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 2046–2065. https://doi.org/10.18653/v1/2020.emnlp-main.161

[10] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[12] Raksha Ramesh, Vishal Anand, Ziyin Wang, Tianle Zhu, Wenfeng Lyu, Serena Yuan, and Ching-Yung Lin. 2020. Kinetics and Scene Features for Intent Detection. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 135–139. https://doi.org/10.1145/3395035.3425641

[13] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., Vancouver, Canada. https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

[14] Fan Yu, DanDan Wang, Beibei Zhang, and Tongwei Ren. 2020. Deep Relationship Analysis in Video with Multimodal Feature Fusion. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 4640–4644. https://doi.org/10.1145/3394171.3416303