Story Semantic Relationships from Multimodal Cognitions

Vishal Anand [1], Raksha Ramesh [1], Ziyin Wang [1], Yijing Feng [2], Jiana Feng [1], Wenfeng Lyu [1], Tianle Zhu [1], Serena Yuan [1], and Ching-Yung Lin [1,2]
[1] {va2361, rn2486, zw2605, jf3283, wl2733, tz2434, sy2657, c.lin}@columbia.edu; [2] {yijing, cylin}@graphen.ai
[1] Columbia University, New York, NY, USA
[2] Graphen AI, New York, NY, USA

ABSTRACT

We consider the problem of building semantic relationship of unseen entities from free-form multi-modal sources. This intelligent agent understands semantic properties by creating (1) logical segments from sources, (2) finds interacting objects, (3) infers their interaction actions using (4) extracted textual, auditory, visual, and tonal information. The conversational dialogue discourses are automatically mapped to interacting co-located objects, and fused with their Kinetic action embeddings at each scene of occurrence. This generates a combined probability distribution representation for interacting entities spanning over every semantic relation class. Using these probabilities, we create knowledge graphs capable of answering semantic queries and infer missing properties in a given context.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Information extraction**; **Lexical semantics**; **Activity recognition and understanding**; *Discourse, dialogue and pragmatics*; *Knowledge representation and reasoning*; *Scene understanding*.

KEYWORDS

natural language processing, information extraction, lexical semantics, video understanding, speaker identification, video to text

ACM Reference Format:
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 (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3394171.3416305

1 INTRODUCTION

With the growing popularity of common sense inference, deep video understanding aims to to automatically deduce relationships between entities in long duration multi-modal inputs and extract knowledge to address varied query-types. With the surge of breakthroughs in text models [5], many tasks have started leveraging transfer-learning on textual data, and recent works now leverage video data, such as event recognition, object detection, and activity recognition. However, a significant knowledge gap exists between joint inference of multiple aspects of the video properties like audio, transcribed conversation and scenes. Key to this intelligent agent is the isolation, fusion and analysis of multimodal features with sufficient annotations for training robust models. In our work, we take a holistic perspective to address the semantic understanding problem in consideration of all available modalities to infer hidden information, and eventually construct knowledge graphs. By fusing different modalities, we gain better understanding of entities, relations and events within movies. In particular, we focus on incorporating reasoning from text and scene, the methodology used including conversational dialog discourse, shot and scene detection, object detection and mapping, and face detection.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3416305

2 BACKGROUND AND RELATED WORK

Many approaches for video understanding adopt the question answering prototype for evaluation. Knowledge graphs are known for capturing both concepts and their pairwise relationships, and their application have been successful to machine learning applications including Web search and social media [6]. Knowledge Graph construction considers three generalized tasks: 1) knowledge extraction, 2) entity mapping, and 3) data integration. Work on multimodal approaches involving knowledge graphs [8] bridges the gap in existing state-of-the-art approaches as it unifies knowledge graphs and deep neural networks in a novel end-to-end learning framework by incorporating external knowledge into video classification. While we propose a multi-modal formulation, [8] has a single-modal approach that takes unstructured text as input and creates a Knowledge Graph with 5 components (Entity Mapping, Co-reference Resolution, Triple Extraction, Triple Integration, and Predicate Mapping). In the domain of social relationship understanding, most existing studies focus on modelling relationships in still images using coarse to fine hierarchical categories [16]. [10] adopts a dual-glance model to make a coarse prediction from objects and appearances while the second glance use contextual cues. [12] propose multi-scale spatiotemporal reasoning framework to capture visual relations between entities. The use of multimodal frameworks in the context of relationships and interaction predictions are not widely explored, which we attempt to address.

3 METHODOLOGY

We place higher importance on context over interactions of individual entities in any of the multi-modal sources. Firstly, we divide resource-rich video files to identify shots that are representative of change of reference of vision. Second, we find a contiguous

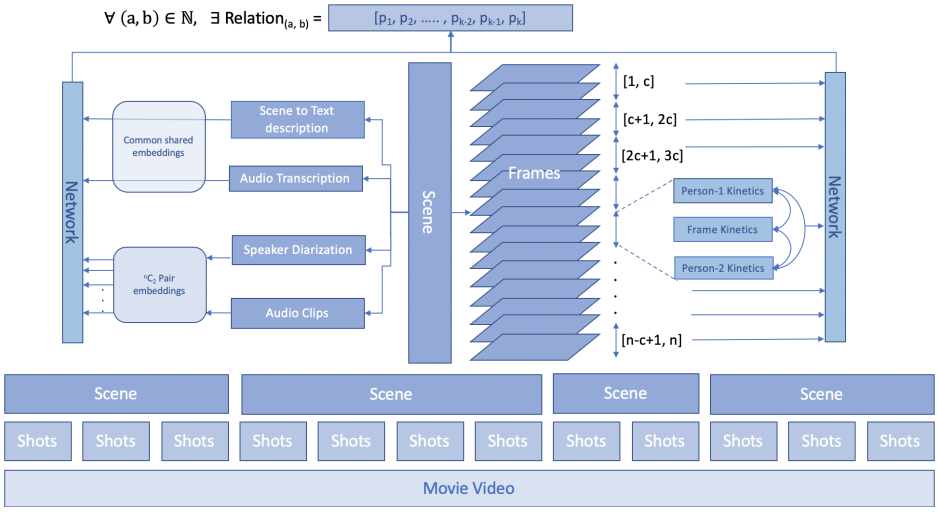

Figure 1: Architecture Schema for multi-modal training.

narrative among the shot-segments, i.e., change in storyline is determined based on when scenes change. This is captured by change in video based light features and maximizing sum of shot-to-shot distance products. With scenes identified as contiguous sets of storylines, we extract their video's textual description, perform audio transcription from the extracted audio, and find diarized texts from individual entities in the scene.

Next, we extract faces in each frame, cluster them and associate with provided ground truth for each movie, and in parallel, find each face's body frames. For each frame, we also extract common objects occurring in the scene to enrich data. Each of these entities are then fed into a kinetics model which emits their probable action being performed. All of these are then used to train a unified model which associates the actions and co-location of entities to produce a set of probabilities for each pair of entities. We also have human evaluation to contrast the performance of our model and add comments on hardness of the problem.

4 DATASET

The HLVU dataset (Table 1) has 10 open source movies sampled from paper [4]. The training set includes four long and two short movies, while testing set includes two long and two short movies. The dataset is annotated with relations between key characters, locations, objects, action events, along with names and images of key entities. The objective is to learn 120 semantic relations between the entities in the dataset using multimodal inputs.

5 BUILDING BLOCKS

5.1 Shot & Scene detection

We perform shot detection on all movies and identify key frames. The frames features extracted are grouped together by similarity scores using sliding-windows into scenes as illustrated in Fig. 2. A scene consists of a sequence of adjacent shots that are semantically related and represent a story within a movie. We cluster multiple shots together and uses some shot-embeddings to deduce if they

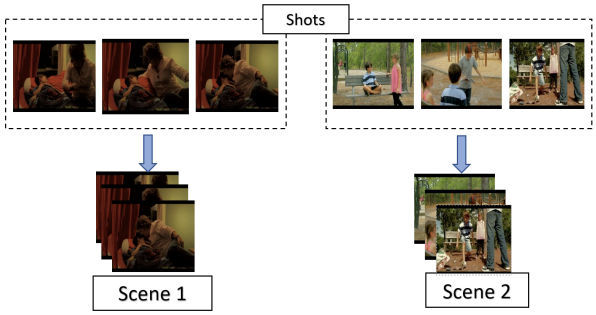

Figure 2: Scene Detection Process

are temporally aligned. We analyze and extend the scene detection work by [13] and adopt the dynamic programming approach to minimize the normalized cost function to group shots together, however we found that using HSV color histograms for features to predict scene boundary accuracies were far more superior than using deep features for movies.

5.2 Object detection and mapping

The entities provided in the HLVU dataset [4] include person, objects, locations and concepts for which relevant images are provided for mapping. The locations and object entities are localized within scenes using SIFT based feature matching to handle varying scales and crops. Keypoints within a frame are computed and descriptors are extracted from each keypoint. Euclidean distance metric is used to compute the best matches between a template and a frame.

5.3 Face detection

The dlib's [9] face detector identifies faces in each video frame. The face detector uses Histogram of Oriented Gradients (HOG) feature combined with a linear classifier. With an image pyramid, and sliding window detection scheme, the detector is able to detect

faces at various scales and locations. These faces are aligned by the 5 facial landmarks detected by dlib's facial landmark predictor.

5.4 Face-Entity Mapping

The aligned faces are mapped to 128D vectors using dlib's ResNet model trained with triplet loss [15]. We also compute the face embeddings for profile images of important characters. We identify the faces by comparing it with the known faces. The unknown faces are first clustered using Chinese whispers [1]. Then, we predict the most similar person with K-nearest neighbors (KNN) for each face detected in the video. The face clusters are labeled with the majority label predicted by KNN. The face clusters with too few entities are labeled as 'unknown'.

5.5 Scene description

We generate scene descriptions at shot-level based on [18] Frame-wise RGB and optical flow features are pooled with region proposals extracted from Faster R-CNN detector. The network is trained on ActivityNet [2] and we observed that the descriptions generated do not accurately capture background objects and scenes. This suggests the need to fine-tune an off-the shelf object detector on a benchmark movie dataset to generate richer descriptions in movies.

5.6 Audio transcription

We use CMU-Sphinx to extract text corresponding to all sound utterances by the cast of given movies. From the initial results, we cross-verify with Google API for each sound utterance's time-stamp. The quality of either of the processes were not very satisfactory.

5.7 Audio emotion embedding

Acoustic features like pitch, energy frequency and spectral coefficients like Mel frequency cepstral coefficients (MFCC) are known to be crucial for emotion recognition and are under-explored in social relationship understanding. We use audio clips for each speaker on scene-levels and use VGG-ish model [7] to extract 128-dimensional semantically compact representation for each second of audio. These are used in downstream relation-classification task. By using labelled relationships from training data with audio embedding, we train a supervised GRU network model.

5.8 Speaker diarization & Face mapping

We extend the CMU-Sphinx to produce speaker-separated audio splits with time-stamps to help with speaker-identification but the output was not satisfactory for the movies. Later we use Google-API with speaker identification and found the results similar on our larger duration audio files. Using the output of both of these systems, we use scene based face-mapping to assign speakers with names by using dlib's 68-facial landmark predictors to capture the shape of lips and estimate lip motion according to relative vertical shape change with respect to face size. We maintain a running average over frames to predict if the target person is the speaker.

5.9 Knowledge Graph relational queries

We construct the knowledge graph to represent what the system has learnt from the movie, where vertices represent entities including people, location, concept and organization, and edges represent relations between two entities. The vertices and edges, together with the confidence of relation prediction are ingested to the graph database on Graphen's Ardi Platform. We also use the Graph Analytic module of Ardi to traverse the graph, retrieve relations given a set of conditions, and get all possible paths between two entities.

6 EXPERIMENTAL-SETUP

6.1 Modality: Scene & Kinetics

To predict a relationship between a pair of entities, we first co-locate and extract individual scene-level tracks for all entities that occur in a scene. A track contains cropped frames of the people and object entities and is temporally aligned with the actual scene. Since clothing and activity are important semantic attributes contributing to model social relationships in videos [16], we extract bounding boxes for a character's body regions using Single Shot Detection (SSD) [11]. The bounding boxes for the character's body regions are localized based on the maximum intersection over union (IoU) with the recognized faces. The scene model uses three parallel video streams - tracks for the pair of entities and the scene as a whole. Features are extracted from the I3D spatio-temporal convolutional network [3] used for activity recognition. We experiment with different durations of video clips to co-locate entities in a scene and found that choosing a 300 frame margin was optimal. The features extracted from the video streams are concatenated and fed to a three-layer MLP trained to predict the sixty relationship categories.

6.2 Modality: Text

Our Text model extends DialogRE [17], based on BERT and we extract relationships between speakers from dialogues. We fine-tune the model and extend their smaller set of relations to 60 relational categories with a fully connected layers. We use our speaker diarization input at scene levels and use data-augmentation by automated mapping of 36 relation-categories to our 60 relational categories by finding neighbors through their embeddings.

7 EVALUATION AND RESULTS

The ACM Grand challenge is based on three different question types on the HLVU dataset [4] - Type 1 requires us to find all valid paths from a given source to the target, for which only one correct solution exists. Therefore F1 scores are chosen as the evaluation metric. A path is considered to be correct only if all the relations and entities along the path match the ground truth. Type 2 is *Fill in the graph space*, where a list of entities and their relations to an unknown entity is given. The answer to this question is a list of potential entities in descending order of prediction confidence. There are totally 60 symmetric relationships to infer from and the Mean Reciprocal Rank (MRR) is the evaluation metric. Type 3, *Multiple choice question answering* is evaluated by accuracy. [1]The Mean Reciprocal Rank allows to capture more retrieval information than F1 and accuracy metrics for Type 1 and Type 3 respectively, which require single solutions.

[1]For test movies' F-1 scores in Table 1, we use Path-1 F1 values. For each movie in training set, we learn from each modality on 5 other train-movies and evaluate on the remaining one. For each test-movie, we train from all movies in training dataset

	Statistics				Evaluation								
					Text			Text+Scene			Human		
	#Actor	#Speaker	#Object	Time	Type1	Type2	Type3	Type1	Type2	Type3	Type1	Type2	Type3
Honey	10	10	12	86 min	0	41.7	0.1	25.0	16.7	0	-	-	-
Nuclear Family	4	4	5	28 min	0	37.5	0	0.0	100.0	0	-	-	-
Spiritual Contact	10	10	13	66 min	0	37.5	0	0.0	52.1	0	66.7	100.0	22.2
Super Hero	7	7	12	18 min	0	0.0	0	25.0	16.7	0	-	-	-
Huckleberry Finn	10	10	20	106 min	0	0.0	0	25.0	6.3	0	100.0	87.5	51.8
Valkaama	7	7	13	93 min	0	25.0	0	25.0	58.3	0	100.0	100.0	50.0
Shooters	8	8	11	41 min	-	-	-	1.2	15.9	50.0	-	-	-
Let's Bring Back Sophie	13	13	22	50 min	-	-	-	0.0	16.7	50.0	-	-	-
The Big Something	9	9	12	101 min	-	-	-	0.0	0.0	50.0	-	-	-
Time Expired	16	16	36	92 min	-	-	-	0.0	0.0	50.0	-	-	-

Table 1: Evaluation in percentages for six Train and four Test movies; Type 1 finds paths between entities (F1), Type 2 fills in missing graph information (MRR), and Type 3 finds semantic relations between nodes (Accuracy)

7.1 Human Ground Truth Evaluation

All training movies are seen by human at least twice and annotated manually in order to create queries and ground-truth for model evaluation[2]. We rarely see performance in the range of 80% MRR for Type 2, and poor results for Type 3 questions, mostly owing to having too many semantic relation classes for a worker to process.

In human evaluation, workers use background music to better understand situation's mood, and grasp conversational context and bodily gestures more readily. Humans have an implicit access to external datasets. Recent approaches focus on data representations but making deductions using multimodal free-form inputs is harder.

7.2 Result Analysis

Our system performs significantly better on question Type 2 - inference task (Table 1). For Type 1, the paths and entities must match the ground truth in order. However, this leads to weaker performance if any entity pairs' relations don't match the ground truth. Since the Text model has soft-transfer learning on pre-trained BERT, we extract a good dialog representation at scene level. For Type 1 questions, the fused model performs better, while the low recall is attributed to information loss due to mismatched/non-identifiable objects from face-object recognition. Imbalanced relationship categories in training data posed a key challenge leading to significant bias. We evaluate the test set using the fused model only.

Type 1 questions were deemed hardest by the challenge authors. Our model performs better on test set when the paths are evaluated on relations grouped into five sets. For *Shooters*, the recall of 2.8% is higher than the precision of 0.075% implying there are fewer false negatives than false positives. We can reduce false positives across movies by encoding features attribute like age, gender, etc. that will help eliminate edges in the knowledge graph between two entities that logically doesn't apply. Low recall is also attributed to:

(1) We do not incorporate inverse relations, i.e. relations outside of the primary sixty categories. The paths between *Robin's father* and *Nicole* in *Let's Bring Back Sophie* contain multiple inverse relations like *Socialized At By*.
(2) When an important entity is not recognized in the pipeline, the body tracks for the kinetics model becomes sparse or non-existent, and the model misses inferring their paths. This is evident in *Shooters* where *Mrs Milton* should connect edges in multiple paths from *Isaac* and *Jaden*.
(3) When two entities are not co-located or are tracked rarely, our model either does not infer a relation between them or does so with a lower confidence, leading to missed edges. For example, most paths between *Emil Oryx* and *Sasha* in *Time Expired* should contain *Corinna Zimmerman*, but is not co-located with entities in the suggested path. One possible way to reduce missed co-located entities is to increase the frame threshold in our model and introduce more hyperparameters.

For Type 2 ranking questions, our model retrieves target entities from the knowledge graph based on the properties of associated edges for multiple queries. However, Type 2 suffers from the same problems described for Type 1. For longer movies, it generates too many sets of probabilities, one for each scene, that hinders entity-pair distribution from converging optimally. This explains the MRR scores for *The Big Something* and *Time Expired*.

In Type 3 questions, our model gets 100% accuracy for questions resembling "How many children/siblings does A have?" across all test movies. These comprises of 50% of the questions in Type 3, therefore we get an accuracy of 50% or more across all movies.

8 CONCLUSION AND FUTURE WORK

We found that segmentation in storylines helped a lot, text-based embeddings are relatively easy to adapt despite the scarcity of training samples in our movie scenes, and audio-emotion embeddings were not effective in detecting critical moments in a movie. Scene based kinetics were very effective in producing large training samples for training and helping create a powerful model. We can improve our character entity co-location pipeline by performing object tracking on each character's occurrence.

A future direction of research on semantic deduction could be based on crisper speaker diarization to prevent garbled transcription, inferring morphological segregation of multilingual conversations [14] and reduce false positive speaker associations using head-tracking as a better proxy for person-face mapping during semantically-deduced sample creation, since freely occurring multimodal data rarely have faces oriented towards the recording device.

[2]The evaluation files on training set are hand-crafted by workers for training movies to best capture performance, as opposed to recording vanilla model loss and accuracies

REFERENCES

[1] Chris Biemann. 2006. Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*. 73–80.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.

[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] 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.

[5] 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

[6] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 601–610.

[7] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[8] Natthawut Kertkeidkachorn and Ryutaro Ichise. 2017. T2KG: An end-to-end system for creating knowledge graph from unstructured text. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

[9] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10 (2009), 1755–1758.

[10] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. 2017. Dual-Glance Model for Deciphering Social Relationships. arXiv:1708.00634 [cs.CV]

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[12] X. Liu, W. Liu, M. Zhang, J. Chen, L. Gao, C. Yan, and T. Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3561–3569.

[13] Daniel Rotman, Dror Porat, Gal Ashour, and Udi Barzelay. 2018. Optimally grouped deep features using normalized cost for video scene detection. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 187–195.

[14] Efsun Sarioglu Kayi, Vishal Anand, and Smaranda Muresan. 2020. MultiSeg: Parallel Data and Subword Information for Learning Bilingual Embeddings in Low Resource Scenarios. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. European Language Resources association, Marseille, France, 97–105. https://www.aclweb.org/anthology/2020.sltu-1.13

[15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[16] Qianru Sun, Bernt Schiele, and Mario Fritz. 2017. A domain based approach to social relation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3481–3490.

[17] Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-Based Relation Extraction. *arXiv preprint arXiv:2004.08056* (2020).

[18] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. 2019. Grounded video description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6578–6587.