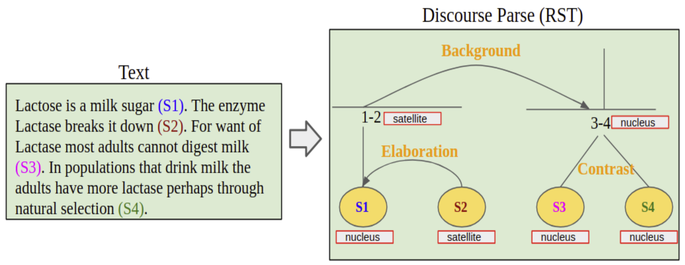
Visual Discourse Parsing

Arjun Akula Song-Chun Zhu

Center for Vision, Cognition, Learning, and Autonomy
University of California, Los Angeles (UCLA)

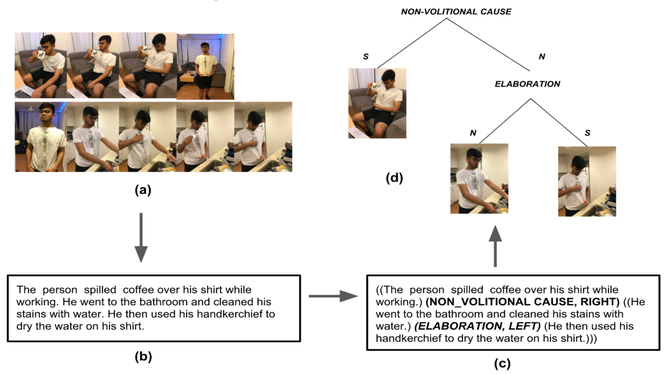
Text-Level Discourse Parsing



Text-level discourse parsing aims to unmask how two segments (or sentences) in the text are related to each other.

Visual Discourse Parsing (VDP)

We introduce a new AI task - **Visual Discourse Parsing**, where the AI agent needs to understand discourse relations among the scenes in a video. Specifically, given a video, the task is to identify a scene's relation with the context.



(a) Video with 9 frames; (b) Textual description of video; (c) RST discourse structure (represented as sequence of words) of description in (b). The notations LEFT and RIGHT represent the direction (nuclearity) of the rhetorical relations; (d) RST discourse structure of the video using 3 frames, i.e. scenes. The symbols N and S indicate the nucleus and satellite of each rhetorical relation.

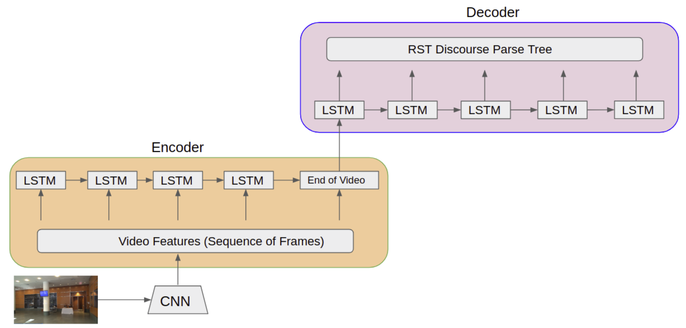
Applications of VDP

- Video Summarization.
- Video Captioning: aids in generating coherent paragraph descriptions of videos.
- Visual Sentiment Analysis.
- Visual Dialog and Visual Story-telling.

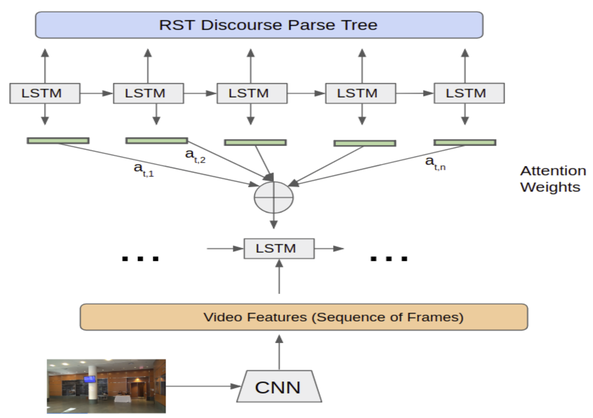
VDP Data Collection

- Developed a VDP dataset containing 310 videos.
- Shot at various places in UCLA such as playing sports (Table Tennis, Frisbee, Tennis, Rugby), bus stop, dining hall, elevator, class-room, library, garden and study room.
- On average, the length of each video is about 19 seconds.
- We first manually generated descriptions of each video and then annotated the discourse structure of these descriptions - with the help of 6 graduate students
- Each video is annotated by at least 2 students.

Baseline Models



(1) **Sequence to Sequence Model**: Encoder-Decoder RNN; feature extraction using VGGNet; sequence of scene features are passed to the encoder; decoder generates the RST parse tree as sequence of words until the end of sentence token is generated.



(2) **Sequence to Sequence with Soft Attention**

Evaluation Metrics

(a) **BLEU score**: to evaluate the translation quality of the discourse structure generated from the videos.

(b) **Relations Accuracy**: total number of relations correctly predicted by the model.

(c) **Edges Accuracy**: total number of edges (i.e. RST node nuclearity directions) correctly predicted by the model.

(d) **Relations+Edges Accuracy**: the predicted discourse structure will be considered correct only if all the relations and the edges are correctly predicted by the model.

Results

RNN Type	*#Hidden Units*	*Bidirectional*	*#Layers*	*Relations*	*Edges*	*Relations+Edges*	*Bleu4*
LSTM	256	NO	1	0.3	0.51	0.21	0.22
LSTM	512	NO	1	0.52	0.62	0.42	0.41
LSTM	1024	YES	1	0.49	0.51	0.42	0.33
LSTM	1024	NO	1	0.35	0.51	0.21	0.34
LSTM	512	NO	2	0.35	0.51	0.21	0.38
LSTM	512	NO	3	0.56	0.62	0.42	0.39
LSTM	512	NO	4	0.56	0.62	0.42	0.39
GRU	512	NO	1	0.3	0.51	0.21	0.33

Evaluation using sequence-to-sequence model without Attention.

RNN Type	*#Hidden Units*	*Bidirectional*	*#Layers*	*#Attention Type*	*Relations*	*Edges*	*Relations+Edges*	*Bleu4*
LSTM	512	NO	1	general	0.63	0.69	0.53	0.59
LSTM	512	NO	1	dot	0.52	0.65	0.45	0.52
LSTM	512	NO	1	concat	0.52	0.65	0.45	0.51
LSTM	512	NO	2	general	0.52	0.65	0.45	0.47
LSTM	512	NO	3	general	0.5	0.65	0.39	0.41

Evaluation using sequence-to-sequence model using Attention.

	Target Prediction	Output Prediction
Okay	<left_config> the person was eating some rice <sequence> <bi_dir> he poured some soy sauce on the rice <sequence> <bi_dir> he continued eating the rice	<left_config> the person eating <sequence> <bi_dir> he then some sauce <sequence> <bi_dir> he continued eating
Okay	<left_config> a person was throwing a frisbee high up <background> <right_dir> he caught it himself <sequence> <bi_dir> he threw it again	<left_config> the person was throwing a frisbee <background> <right_dir> he ran it up a frisbee <sequence> <bi_dir> he threw again
Not bad	<left_config> the person was looking at the corner <background> <right_dir> another person showed up <justify> <right_dir> they greeted each other and sat together	<left_config> the person was throwing a frisbee <background> <right_dir> another person came and a seat and him to the person
worst	<right_config> the person was drinking up his juice <sequence> <bi_dir> he went to get a cup of coffee <elaboration> <left_dir> he put some sugar into it and went back to his seat	<left_config> the person was looking at the computer <background> <right_dir> he put to to to to to to to to to to to

Qualitative Analysis