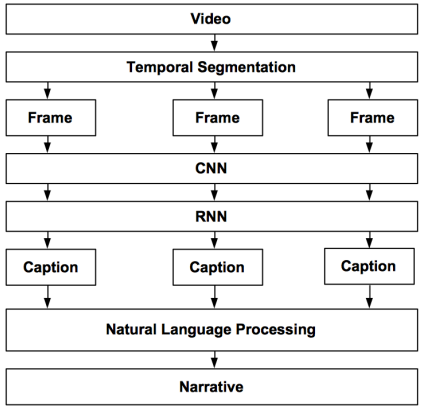
Beyond Caption To Narrative: Video Captioning with Multiple Sentences

Andrew Shin, Katsunori Ohnishi, Tatsuya Harada
The University of Tokyo

Introduction

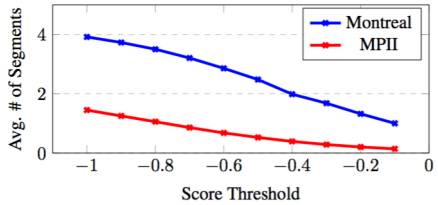
- Most works on video captioning focus on generating single sentence caption
- **GOAL:** to generate captions consisting of multiple sentences from multiple key frames to fully take advantage of richer contents in video
- **Contributions:**
  - key frames are selected from segmenting the video by change of "action"
  - generate caption for each frame, and connect them with natural language processing techniques
- Overall workflow:



Temporal Segmentation

- Goal is to segment the video temporally, not to classify the actions
- Sliding window method and motion features only for temporal segmentation
- iDT to extract motion features
- UCF 101 to learn actions

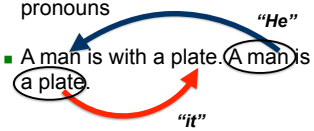
- Temporal sliding windows of length 30,60,90, and 120 frames
- Slide the window in steps of 30 frames
- Re-score detected windows with non-maximum suppression to remove overlapping
- To avoid redundant segmentation, we set a score threshold



- -0.5 for Montreal dataset, -1.0 for MPII for roughly 2 frames per video

Backward Coreference Resolution

- After extracting middle frame from each segment, and generating captions:
- There are no anaphora, but repeated appearances of noun phrases as antecedent candidates
- Want to link the coreferences, then convert the later references to appropriate pronouns
- A man is with a plate. A man is sitting with a plate.



- Assign gender tokens with Stanford CoreNLP for singular human subjects
- Run coreference resolution for non-singular or non-human references

Connective Word Generation

- Want to find appropriate transitions
- Hard to determine the number of classes and collect training data → unsupervised way
- Collected 500 instances of adjacent sentences from Wikicorpus, with connective term in 2nd sentence
- Converted to 300-dim sentence2vec
- Simply match the generated captions with adjacent sentences by L2 distance, and insert the connective word present

Experiment

- 1st baseline: caption from single middle frame
- 2nd baseline: winner of LSMDC challenge
- Extract HOG, HOF, MBH along iDT, and encode with FV
- 1 vs. rest linear SVM with C=100
- Extract middle frame from each segment, and extract fc7 from VGG
- Pass CNN features to LSTM to generate captions, and apply backward coreference resolution and connective word generation

Dataset	Model	BLEU4	CIDEr	METEOR
Montreal	Mid-frame	0.003	0.070	0.042
	Ours	**0.004**	**0.089**	**0.047**
MPII	Mid-frame	0.009	0.065	0.046
	Ours	**0.013**	**0.075**	**0.048**
MS Video	Mid-frame	0.043	0.148	0.107
	Ours	**0.063**	**0.177**	0.104

Model	Avg. Length	BLEU4	CIDEr	METEOR
LSMDC [1]	5.33	**0.006**	0.092	0.058
Mid-frame	9.78	0.004	0.068	0.044
Ours	**19.34**	**0.006**	0.085	0.048

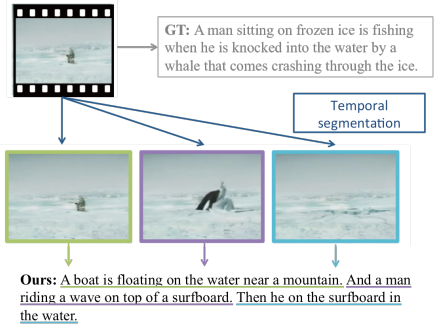
- example from Montreal dataset

A man is holding a box of doughnuts. | A man and a woman are standing next to each other. | A woman is holding a plate of food.

"A man is holding a box of doughnuts. Then he and a woman are standing next to each other. Then she is holding a plate of food."

Ground Truth	***With the table gone, all of the girls sit in their same seats, their trays on their laps.***
Mid-frame	***A man and a woman standing next to each other.***

- Applicable to non-movie clips?
- Applied our method to subset of MS Video Description corpus with score threshold -0.1
- Compared the captions from sub-interval with ground truths



Conclusion

- Proposed a novel method for a story-like video captioning

Reference

- [1] Shetty and Laaksonen "*Video captioning with recurrent networks based on frame and video-level features and visual content classification*" in ICCV Workshop on LSMDC 2015