SODA: Story Oriented Dense Video Captioning Evaluation Flamework

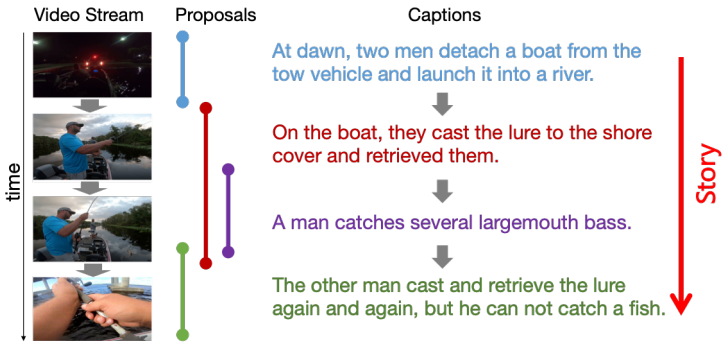
Soichiro Fujita*, Tsutomu Hirao**, Hidetaka Kamigaito*, Manabu Okumura*, Masaaki Nagata**
* Tokyo Institute of Technology , ** NTT Communication Science Laboratories

Source code ->

Background

・Each video has a story
・Current systems sometimes generate too many captions

(https://www.youtube.com/watch?v=d41k1dkdqiY)



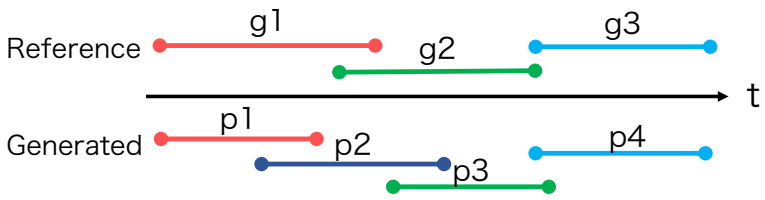
Problems of the current scorer

1. Sometimes giving high scores to redundant captions
2. Disregarding the story (the ordering of captions)

Proposed method : SODA

・1-to-1 matching to maximize temporal overlap via DP
・Deriving F-measure scores from the METEOR scores

Comparison between methods



Tab.1 IoU table

	p1	p2	p3	p4
g1	0.7	0.6	0.0	0.0
g2	0.0	0.5	0.6	0.0
g3	0.0	0.0	0.1	0.9

IoU: temporal overlap between g and p

	Matching Proposals	Caption Evaluation
Current	All pairs which exceeds the threshold (tIoU) (g1, p1), (g1, p2), (g2,p3) (g2, p2), (g3, p4)	Average based on the # of pairs $E(\mathcal{G},\mathcal{P},\tau)=\frac{\sum_{p\in P}\sum_{g\in G_{p,\tau}}\mathrm{METEOR}(g,p)}{\sum_{p\in P}\lvert G_{p,\tau}\rvert}$ # of pairs
SODA (our method)	Dynamic programming (g1, p1),(g2, p3),(g3, p4) Maximize sum(IoU x METEOR)	F-measure $\mathrm{Precision}(\mathcal{G},\mathcal{P})=\frac{\sum_{g\in\mathcal{G}}\mathrm{IoU}\times\mathrm{METEOR}(g,p_{a(g)})}{\lvert\mathcal{P}\rvert}$ $\mathrm{Recall}(\mathcal{G},\mathcal{P})=\frac{\sum_{g\in\mathcal{G}}\mathrm{IoU}\times\mathrm{METEOR}(g,p_{a(g)})}{\lvert\mathcal{G}\rvert}$

Experiment

・Dataset
- ActivityNet Captions
- # of videos: 4883

・Model
- E2E Transformer [Zhou+ 2018]
- # of output captions pre video: 200

• Detecting inappropriate captions

Randomly changed the ratio of generated captions
→ SODA gives **high scores** to **the optimal number of captions**

Methods	m = 0.1	0.5	1.0	2.0	10	All
Current	3.78	4.04	4.10	4.14	4.18	4.19
SODA	1.47	3.41	4.02	3.83	1.70	0.63

• Detecting incorrect order

Randomly changed the order of generated captions
→ SODA is **more sensitive** to **the incorrect order of captions**

Methods	Correct	Swap	Shuffle
Current	16.1	14.5 (-10.2%)	10.8 (-33.1%)
SODA	17.8	14.5 (-18.9%)	7.66 (-57.0%)