SODA: Story Oriented Dense Video Captioning Evaluation Framework

Soichiro Fujita [1], Tsutomu Hirao [2], Hidetaka Kamigaito [1],
Manabu Okumura [1], Masaaki Nagata [2]

1. Tokyo Institute of Technology
2. NTT Communication Science Laboratories

Background

Each video has a story

Video Stream

Proposals

Captions

time

At dawn, two men detach a boat from the tow vehicle and launch it into a river.

On the boat, they cast the lure to the shore cover and retrieved them.

A man catches several largemouth bass.

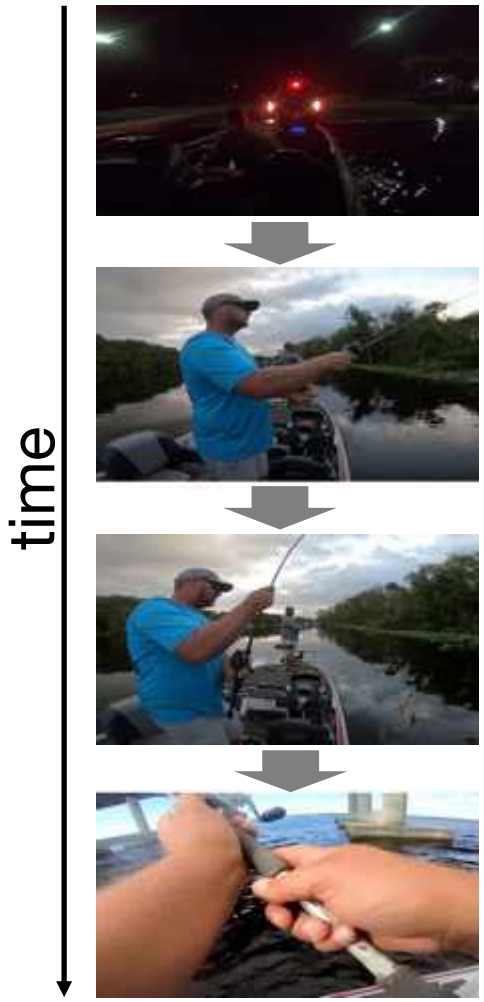
The other man cast and retrieve the lure again and again, but he can not catch a fish.

Story

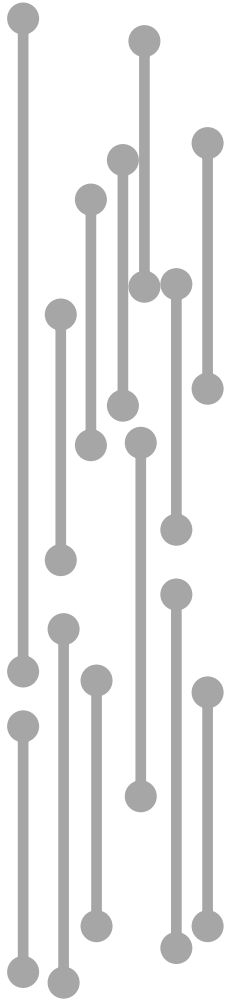
"Tough Day Bass Fishing The St Johns River" by RivweBassin (https://www.youtube.com/watch?v=d41k1dkdqiY)

Current systems sometimes generate too many captions

Video Stream



Proposals



Captions

At dawn, two men detach a boat from the tow vehicle and launch it into a river.

A man catches several largemouth bass.

On the boat, they cast the lure to the shore again and again, but he can not catch a fish.

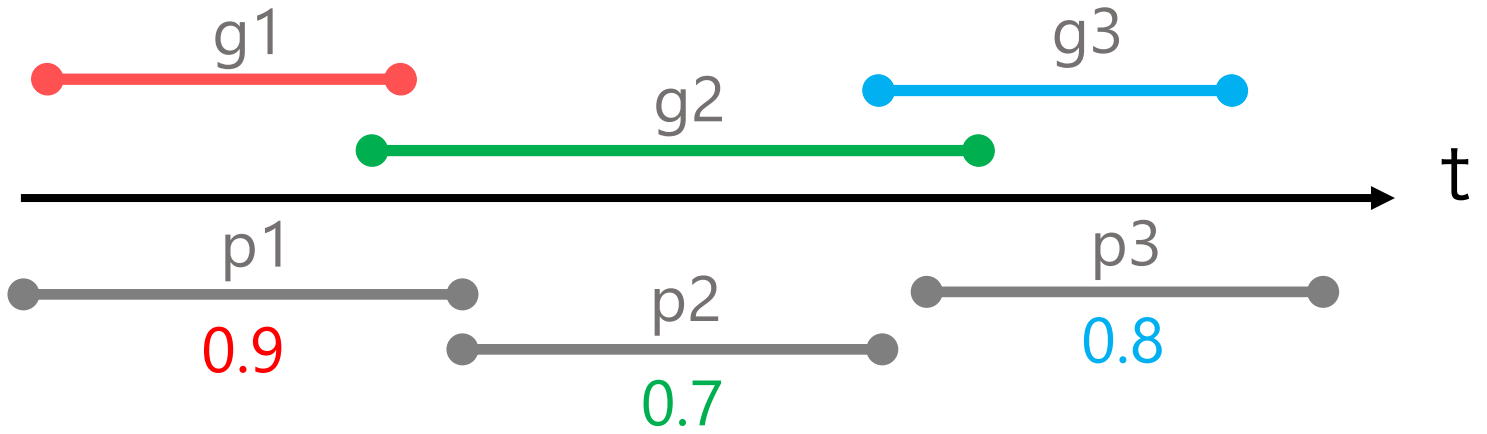
The other man cast and retrieve the lure cover and retrieved them.

"Tough Day Bass Fishing The St Johns River" by RivweBassin (https://www.youtube.com/watch?v=d41k1dkdqiY)

Evaluation for dense video captioning systems

Evaluate two components:

- Temporal localized events

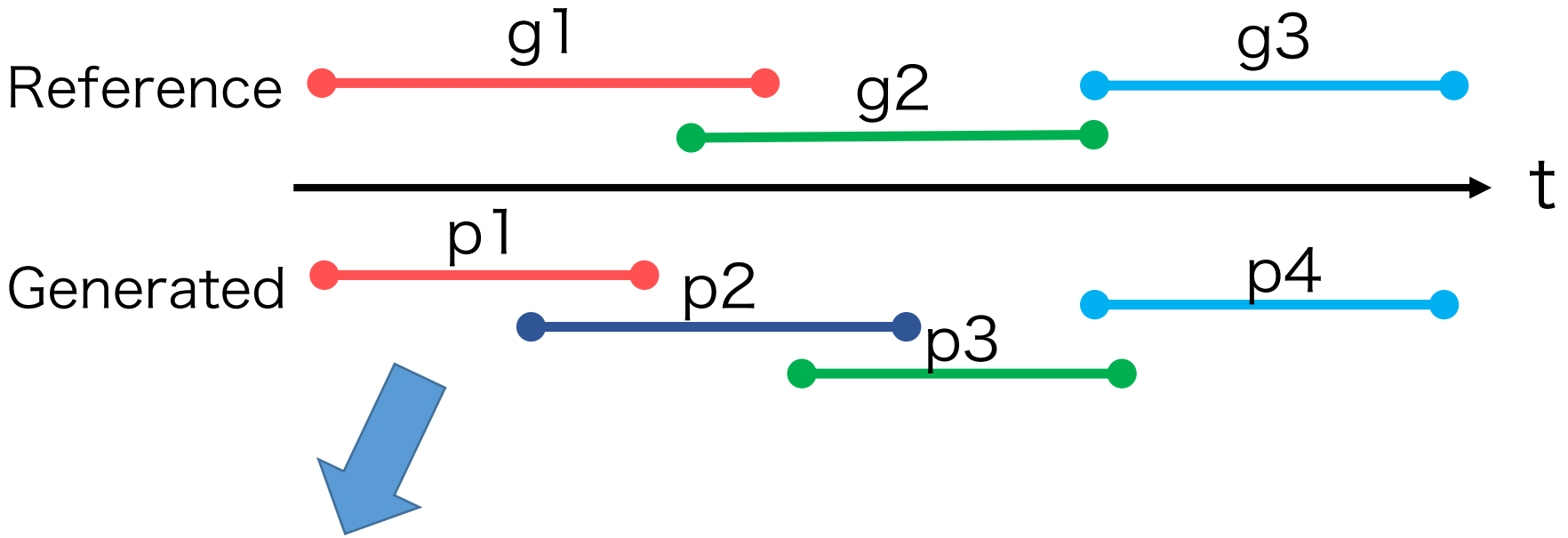


- Generated captions

METEOR(g1, p1), METEOR(g2,p2), METEOR(g3,p3)

Current evaluation framework: ActivityNet Captions scorer

Step1: Calculation of a 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	0.05	0.9

IoU: Intersection over Union

temporal overlap rate of proposal

$$\mathrm{IoU}(g,p)=\max\left(0,\frac{\min(e(g),e(p))-\max(s(g),s(p))}{\max(e(g),e(p))-\min(s(g),s(p))}\right)$$

Step2: Finding pairs and averaging scores

	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	0.05	0.9

IoU Threshold τ : 0.5

Pairs exceeding the threshold:
(g1, p1), (g1,p2), (g2, p2),
(g2, p3), (g3, p4)

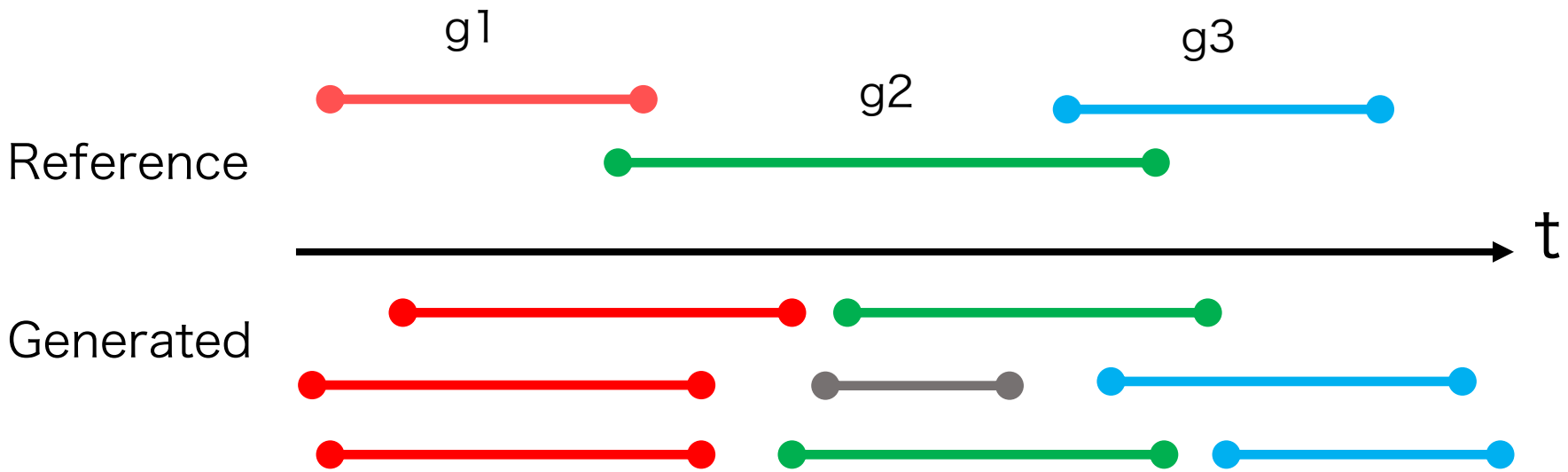
Averaging METEOR scores of all the pairs

$$E(\mathcal{G},\mathcal{P},\tau) = \frac{\sum_{p\in P}\sum_{g\in G_{p,\tau}} \mathrm{METEOR}(g,p)}{\sum_{p\in P}|G_{p,\tau}|}$$

$\sum_{p\in P}|G_{p,\tau}|$: # of matched pairs

Problems of ActivityNet Captions scorer

- Disregarding the story (the ordering of captions)
- Sometimes giving high scores to redundant captions

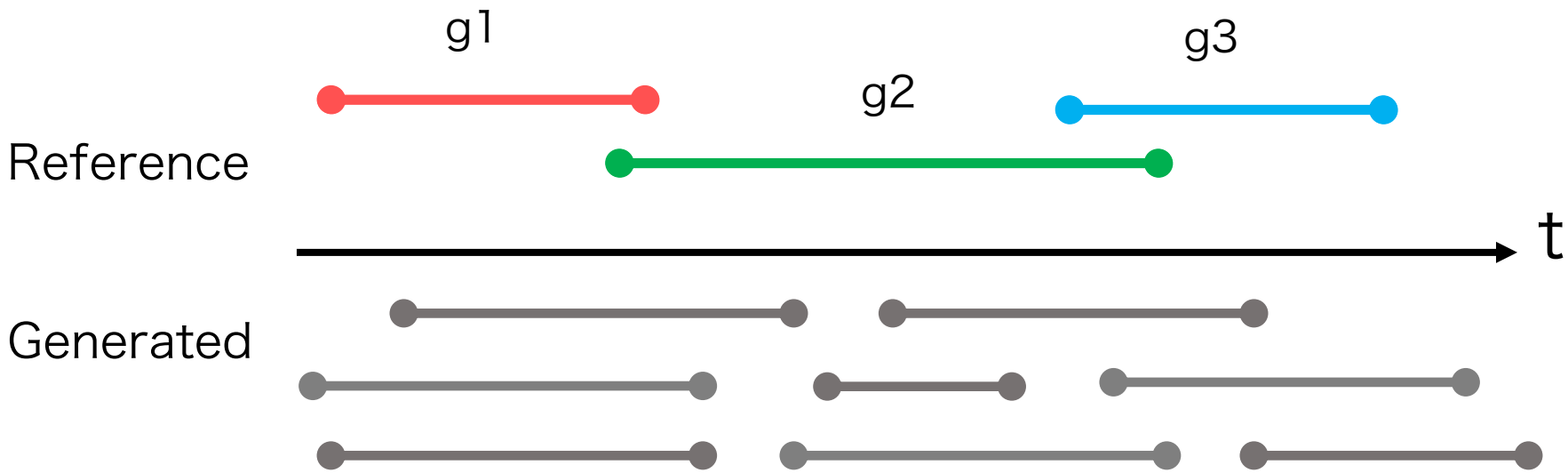


Proposed method: SODA

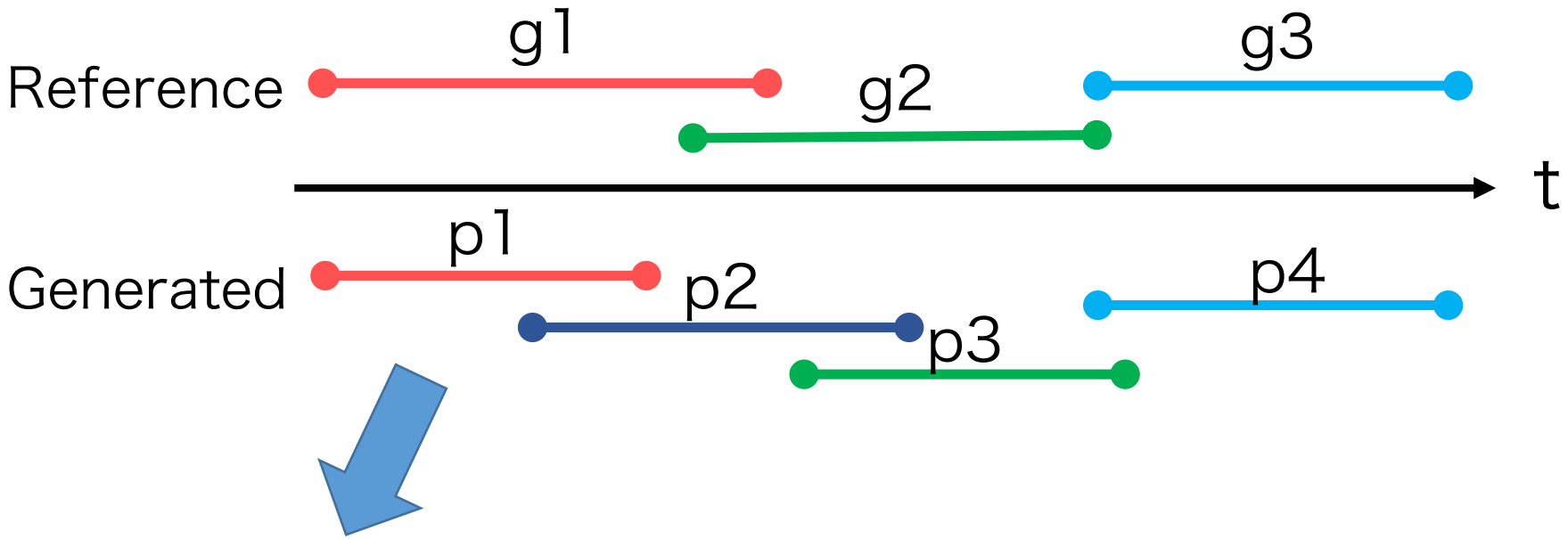
Proposed method: SODA

Advantage:

- Considering the story (the ordering of captions)
- Preventing redundant captions from obtaining good scores



Step1: Calculation of a cost 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	0.05	0.9

Cost: IoU x METEOR

$$C_{i,j} = \mathrm{IoU}(g_i, p_j) f(g_i, p_j).$$

Step2: Filling a DP table

Cost 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	0.05	0.9

Cost: $C_{i,j} = \mathrm{IoU}(g_i, p_j) f(g_i, p_j).$

State: $S[i][j] = \max \begin{cases} S[i-1][j], \\ S[i-1][j-1] + C_{i,j}, \\ S[i][j-1]. \end{cases}$

DP table

	S[*][0]	S[*][1]	S[*][2]	S[*][3]	S[*][4]
S[0][*]	0.0	0.0	0.0	0.0	0.0
S[1][*]	0.0				
S[2][*]	0.0				
S[3][*]	0.0				

Step2: Filling a DP table

Cost 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	0.05	0.9

Cost: $C_{i,j} = \mathrm{IoU}(g_i, p_j) f(g_i, p_j).$

State: $S[i][j] = \max \begin{cases} S[i-1][j], \\ S[i-1][j-1] + C_{i,j}, \\ S[i][j-1]. \end{cases}$

D

	S[*][0]	S[*][1]			
S[0][*]	0.0	0.0			
S[1][*]	0.0	0.7			
S[2][*]	0.0				
S[3][*]	0.0				

$$S[1][1] = max \begin{cases} s[0][1] = 0.0 \\ s[0][0] + C_{1,1} = 0.0 + 0.7 \\ s[1][0] = 0.0 \end{cases}$$

Step2: Filling a DP table

Cost 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	0.05	0.9

Cost: $C_{i,j} = \mathrm{IoU}(g_i, p_j) f(g_i, p_j).$

State: $S[i][j] = \max \begin{cases} S[i-1][j], \\ S[i-1][j-1] + C_{i,j}, \\ S[i][j-1]. \end{cases}$

DP table

	S[*][0]	S[*][1]	S		
S[0][*]	0.0	0.0			
S[1][*]	0.0	0.7	0.7		
S[2][*]	0.0	0.7			
S[3][*]	0.0				

$$S[1][2] = max \begin{cases} s[0][2] = 0.0 \\ s[0][1] + C_{1,2} = 0.0 + 0.6 \\ s[1][1] = 0.7 \end{cases}$$

Step2: Filling a DP table

Cost 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	0.05	0.9

Cost: $C_{i,j} = \mathrm{IoU}(g_i, p_j) f(g_i, p_j).$

State: $S[i][j] = \max \begin{cases} S[i-1][j], \\ S[i-1][j-1] + C_{i,j}, \\ S[i][j-1]. \end{cases}$

DP

	S[*][0]	S[*][1]			
S[0][*]	0.0	0.0			
S[1][*]	0.0	0.7	0.7	0.7	
S[2][*]	0.0	0.7	1.2		
S[3][*]	0.0	0.7			

$$S[2][2] = max \begin{cases} s[1][2] = 0.7 \\ s[1][1] + C_{2,2} = 0.7 + 0.5 \\ s[2][1] = 0.7 \end{cases}$$

Step2: Filling a DP table

Cost 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	0.05	0.9

Cost: $C_{i,j} = \mathrm{IoU}(g_i, p_j) f(g_i, p_j).$

State: $S[i][j] = \max \begin{cases} S[i-1][j], \\ S[i-1][j-1] + C_{i,j}, \\ S[i][j-1]. \end{cases}$

$S[2][3] = max \begin{cases} s[1][3] = 0.7 \\ s[1][2] + C_{2,3} = 0.7 + 0.6 \\ s[2][2] = 1.2 \end{cases}$

	S[*][0]	S[*][1			
S[0][*]	0.0	0.0			
S[1][*]	0.0	0.7	0.7	0.7	0.7
S[2][*]	0.0	0.7	1.2	1.3	
S[3][*]	0.0	0.7	1.2		

Step2: Filling a DP table

Cost 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	0.05	0.9

Cost: $C_{i,j} = \mathrm{IoU}(g_i, p_j) f(g_i, p_j).$

State: $S[i][j] = \max \begin{cases} S[i-1][j], \\ S[i-1][j-1] + C_{i,j}, \\ S[i][j-1]. \end{cases}$

DP

	S[*][0]	S[*][1]			
S[0][*]	0.0	0.0			
S[1][*]	0.0	0.7	0.7		0.7
S[2][*]	0.0	0.7	1.2	1.3	1.3
S[3][*]	0.0	0.7	1.2	1.3	

$$S[3][3] = max \begin{cases} s[2][3] = 1.3 \\ s[2][2] + C_{3,3} = 1.2 + 0.05 \\ s[3][2] = 1.2 \end{cases}$$

Step2: Filling a DP table

Cost 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	0.05	0.9

Cost: $C_{i,j} = \mathrm{IoU}(g_i, p_j) f(g_i, p_j).$

State: $S[i][j] = \max \begin{cases} S[i-1][j], \\ S[i-1][j-1] + C_{i,j}, \\ S[i][j-1]. \end{cases}$

DP

	S[*][0]	S[*][1]			
S[0][*]	0.0	0.0			
S[1][*]	0.0	0.7	0.7		0.7
S[2][*]	0.0	0.7	1.2	1.3	1.3
S[3][*]	0.0	0.7	1.2	1.3	2.2

$$S[3][4] = max \begin{cases} s[2][4] = 1.3 \\ s[2][3] + C_{3,4} = 1.3 + 0.9 \\ s[3][3] = 1.3 \end{cases}$$

Step3: Tracing back to the routes and finding the pairs

Cost 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	0.05	0.9

Cost: $C_{i,j} = \mathrm{IoU}(g_i, p_j) f(g_i, p_j).$

State: $$S[i][j] = \max \begin{cases} S[i-1][j], \\ S[i-1][j-1] + C_{i,j}, \\ S[i][j-1]. \end{cases}$$

DP table

	S[*][0]	S[*][1]	S[*][2]	S[*][3]	S[*][4]
S[0][*]	0.0	0.0	0.0	0.0	0.0
S[1][*]	0.0	**0.7**	0.7	0.7	0.7
S[2][*]	0.0	0.7	1.2	**1.3**	1.3
S[3][*]	0.0	0.7	1.2	1.3	**2.2**

Pairs

(g1, p1)
(g2, p3)
(g3, p4)

Step4: Calculation of F-measure

	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	0.05	0.9

Pairs: (g1, p1), (g2, p3), (g3, p4)

Sum of scosts: 2.2

$$\text{Precision}(\mathcal{G},\mathcal{P}) = \frac{\sum_{g\in\mathcal{G}} \text{IoU} \times \text{METEOR}(g, p_{a(g)})}{|\mathcal{P}|} = \frac{2.2}{4} = 0.55$$

$$\text{Recall}(\mathcal{G},\mathcal{P}) = \frac{\sum_{g\in\mathcal{G}} \text{IoU} \times \text{METEOR}(g, p_{a(g)})}{|\mathcal{G}|} = \frac{2.2}{3} = 0.73$$

$$\text{F-measure}(\mathcal{G},\mathcal{P}) = \frac{2 \times \text{Precision}(\mathcal{G},\mathcal{P}) \times \text{Recall}(\mathcal{G},\mathcal{P})}{\text{Precision}(\mathcal{G},\mathcal{P}) + \text{Recall}(\mathcal{G},\mathcal{P})} = \mathbf{0.63}$$

Experiments

Experimental settings

Dataset: ActivityNet Captions Dataset

Train 10,024 videos, Validation 4,915 videos

(Test set is not available)

Models:

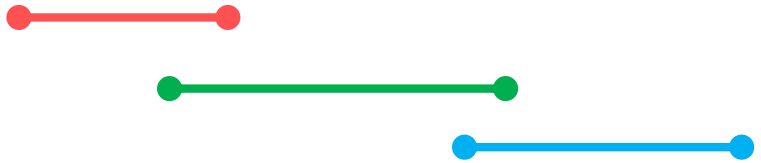
- **E2E Transformer [Zhou+ 2018]**
  The number of output captions per video is **228.21** on average
- **LSTM [Wang+ 2018]**
  The number of output captions per video is **97.10** on average

Exp1: Detecting inappropriate captions

Does SODA give low scores to too many or too few captions?

- Randomly changed $\boldsymbol{m}$: the ratio of generated captions to a reference

Reference



$0 < \boldsymbol{m} < 1$

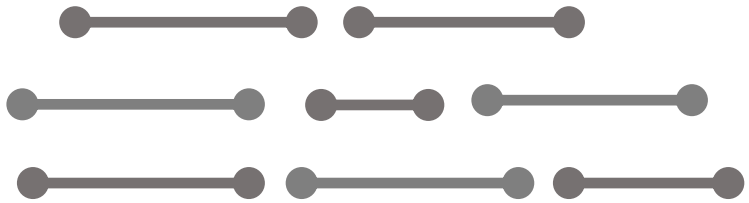
Too few captions

$\boldsymbol{m} = 1$

The same # of captions

$\boldsymbol{m} > 1$

Too many captions



Exp1 results of E2E Transformer

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	Precision	5.93	4.89	4.02	2.87	0.94	0.33
	Recall	0.84	2.61	4.02	5.74	9.30	10.62
	F1	1.47	3.41	4.02	3.83	1.70	0.63

Change of scores (%) when varying the number of captions.

Exp1 results of E2E Transformer

The current method gives **similar scores** regardless number of captions

Methods		0.1	0.5	1.0	2.0	10	All
Current		3.78	4.04	4.10	4.14	4.18	4.19
SODA	Precision	5.93	4.89	4.02	2.87	0.94	0.33
	Recall	0.84	2.61	4.02	5.74	9.30	10.62
	F1	1.47	3.41	4.02	3.83	1.70	0.63

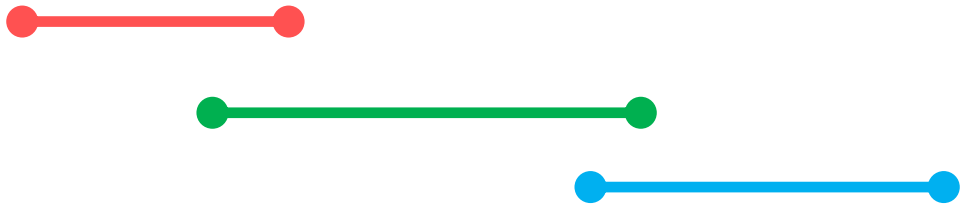
Change of scores (%) when varying the number of captions.

SODA gives **high scores** to **the optimal number of captions**

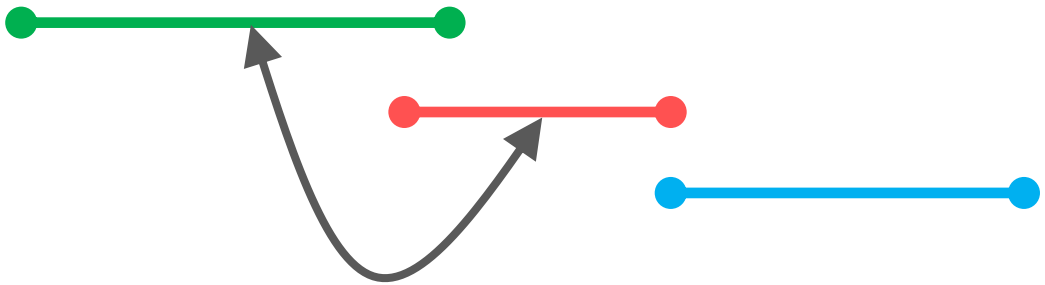
Exp2: Detecting incorrect ordering

Does SODA give low scores to captions with incorrect order?

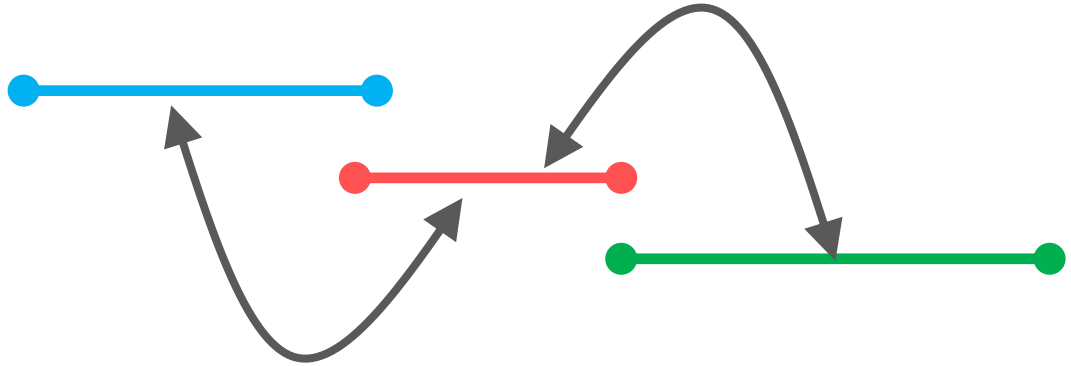
- Correct order of oracle captions:



- Two variations of incorrect order:
  - **Swap**



  - **Shuffle**



Exp2 results of E2E Transformer

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%)

Evaluation scores (%) for oracle captions with correct and incorrect order

The percentage decrease from the score for Correct

Exp2 results of E2E Transformer

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%)

Evaluation scores (%) for oracle captions with cor and incorrect order
The percentage decrease from the sco rrect

SODA is **more sensitive** to
the incorrect order of captions

Manual evaluation: pairwise comparison




Caption A

At dawn, two men detach a boat from the tow vehicle and launch it into a river.

On the boat, they cast the lure to the shore cover and retrieved them.

A man catches several largemouth bass.

The other man cast and retrieve the lure again and again, but he can not catch a fish.

VS.

Caption B

On the boat, they cast the lure to the shore cover and retrieved them.

At dawn, two men detach a boat from the tow vehicle and launch it into a river.

The other man cast and retrieve the lure again and again, but he can not catch a fish.

A man catches several largemouth bass.

☑ ☐ ☐ ☐ ☐

A is better B is better

50 videos x 10 people

"Tough Day Bass Fishing The St Johns River" by RivweBassin (https://www.youtube.com/watch?v=d41k1dkdqiY)

1. Comparing the performance of systems

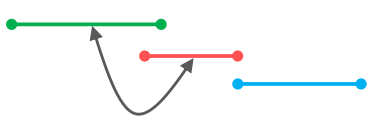
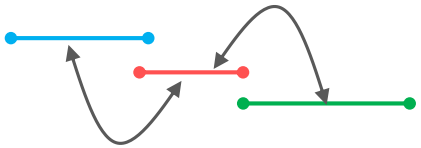
	Caption A		Caption B
	E2E Transformer oracles	VS.	LSTM oracles
Human judgment:	80%		20%

Accuracy of frameworks against the human judgment:

- Current: **0.66**
- SODA: **0.76**

SODA can evaluate systems more consistent with human judgment

2. Comparing the incorrect order of captions

	Caption A		Caption B
	Swapped captions	VS.	Shuffled captions 
Human judgment:	94%		6%

Accuracy of frameworks against the human judgment:

- Current: 0.72
- SODA: 0.94

SODA gives a penalty more consistent with human judgment

Conclusion

Conclusion

SODA:

- A new evaluation framework for dense video captioning

Architecture:

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

source code

Result:

- Considering the story (the ordering of captions)
- Preventing redundant captions from obtaining good scores
- Being consistent to the manual evaluations