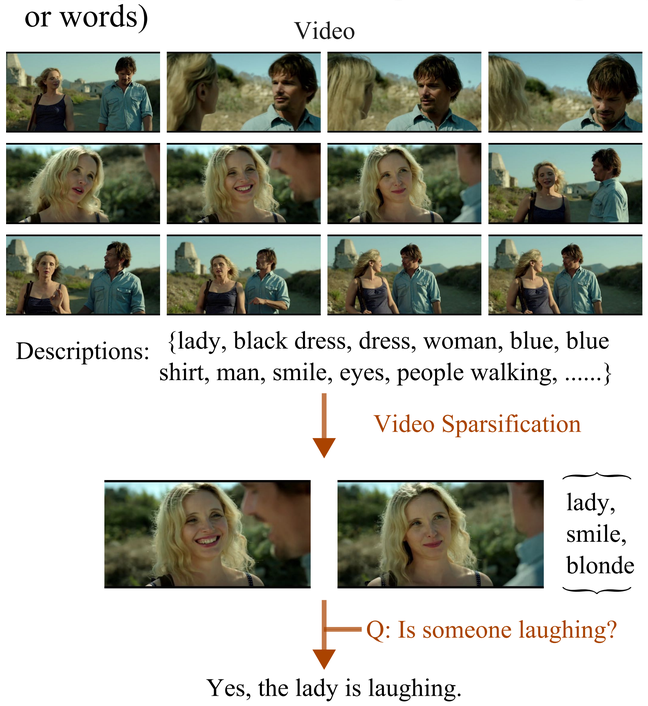
Charactering Video Question Answering with Sparsified Inputs

Shiyuan Huang[1], Robinson Piramuthu[2], Vicente Ordonez[2,3], Shih-Fu Chang[1], Gunnar A. Sigurdsson[2]
[1]Columbia University, [2]Amazon Alexa AI, [3]Rice University

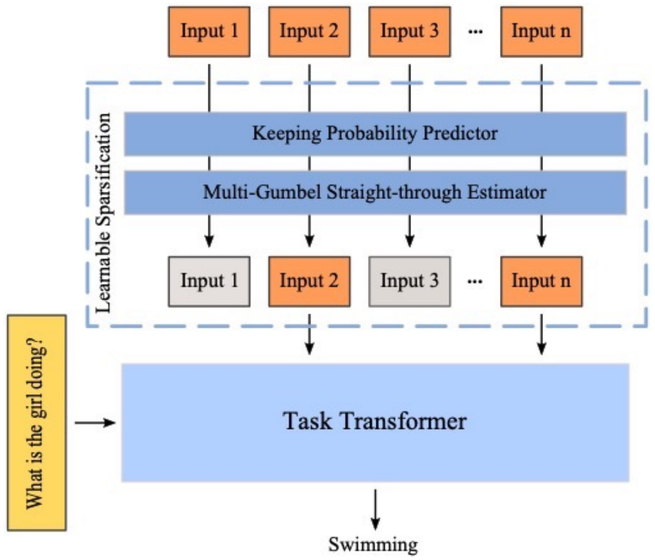
Sparsified VideoQA

- Accomplish VideoQA with sparse video inputs (frames or words)



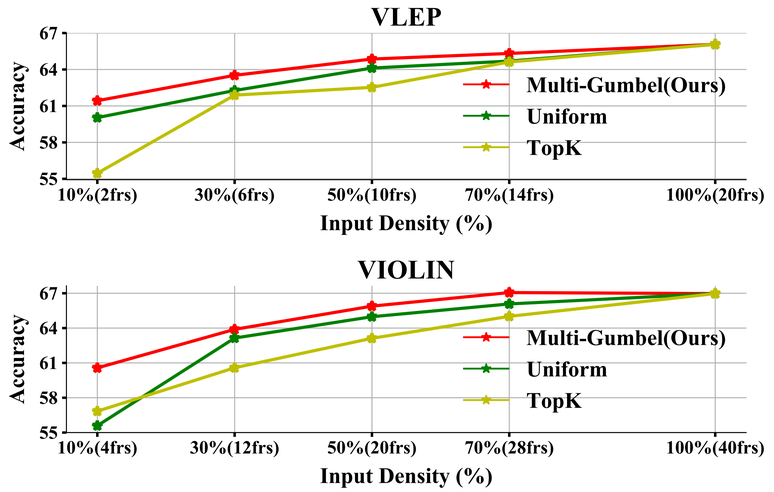
Learnable Sparsification

- Learn to keep important tokens
- Multi-Gumbel Straight-through Estimator: facilitate sampling process during training



VideoQA Results

- About 5% performance drop with 10% video lengths



Multi-modal VideoQA Results

		Visual (Snippets)				
		0	1 snippet	2 snippets	5 snippets	20 snippets
Textual (Words)	0	14.6 (Q-only)	28.65	30.24	31.26	35.43 [22]
	5 words	17.5	28.68	30.31	31.70	35.43
	10 words	18.22	29.87	31.43	31.88	36.01
	25 words	20.14	30.16	31.59	32.03	36.09
	100 words	26.75	31.47	32.11	33.21	**36.42**

Ablation on Multi-Gumbel Process

- More explorative selection is beneficial for denser inputs

	VLEP			VIOLIN		
Input Percentage	$\tau = 0.01$	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.01$	$\tau = 0.1$	$\tau = 0.5$
10%	**60.25**	56.01	58.94	56.25	**60.57**	58.80
30%	60.95	**63.52**	59.13	61.72	57.64	**62.34**
50%	63.05	63.64	**64.30**	65.57	64.48	**66.06**
70%	63.73	65.14	**65.32**	65.94	66.52	**67.06**

Sparsified Video Examples