SEMANTIC-PRESERVING METRIC LEARNING FOR VIDEO-TEXT RETRIEVAL

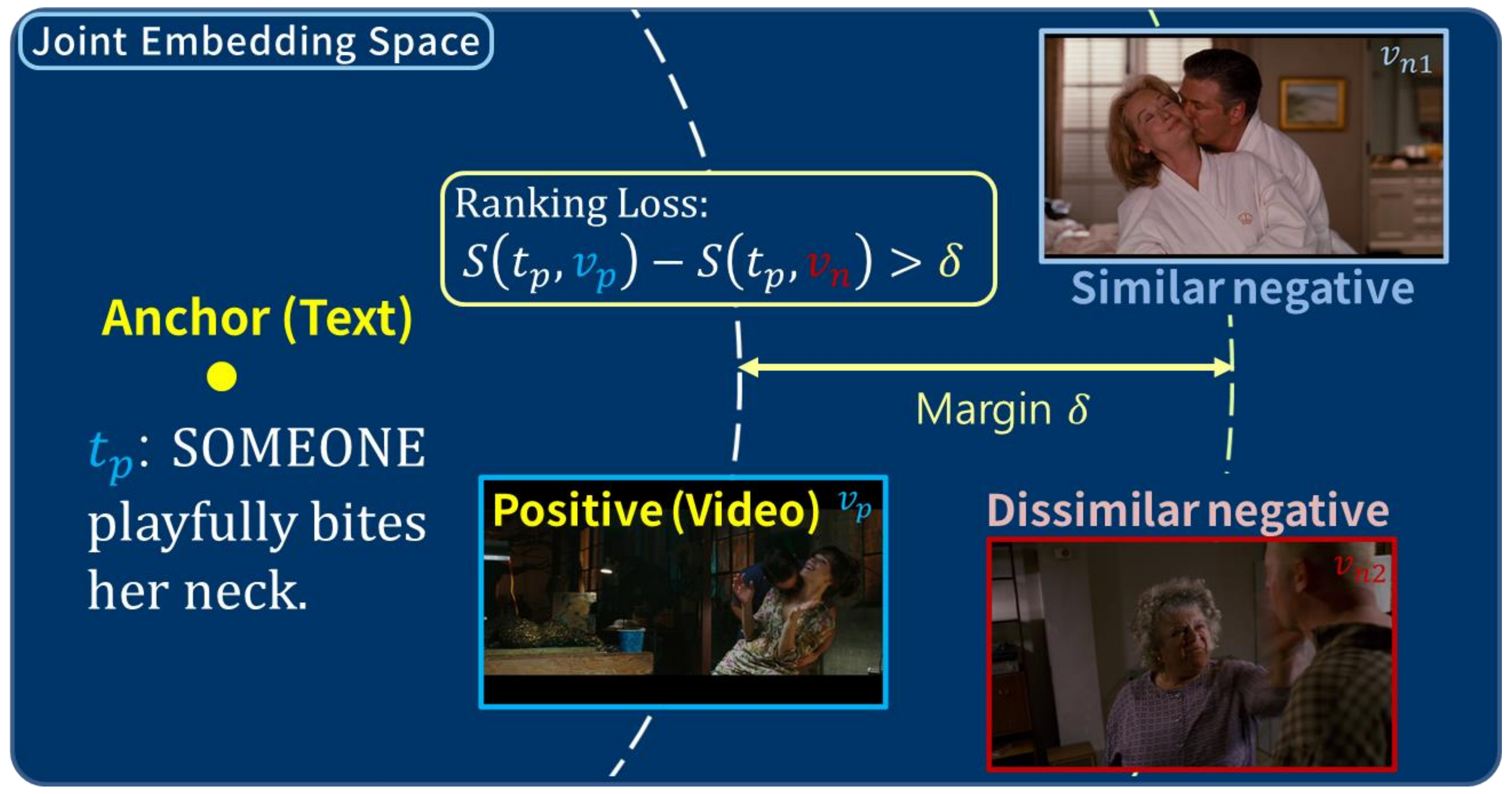
Paper #1006

Sungkwon Choo, Seong Jong Ha, Joonsoo Lee

Vision AI Lab, AI Center, NCSOFT

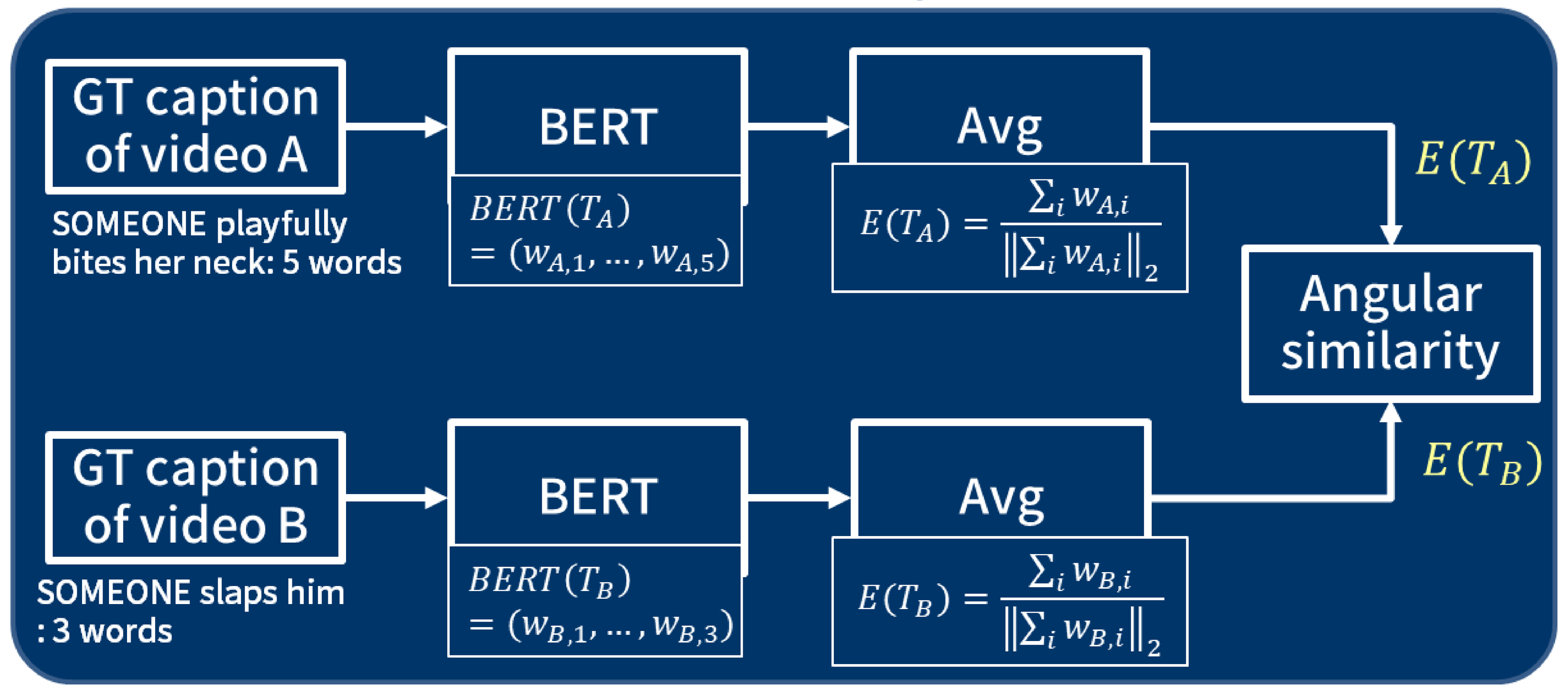
Motivations

Ranking loss treats all negatives "equally"
→ Resulting in a large semantic discrepancy



Semantic relevance

Define a measure of non-binary semantic relations of corresponding GT captions

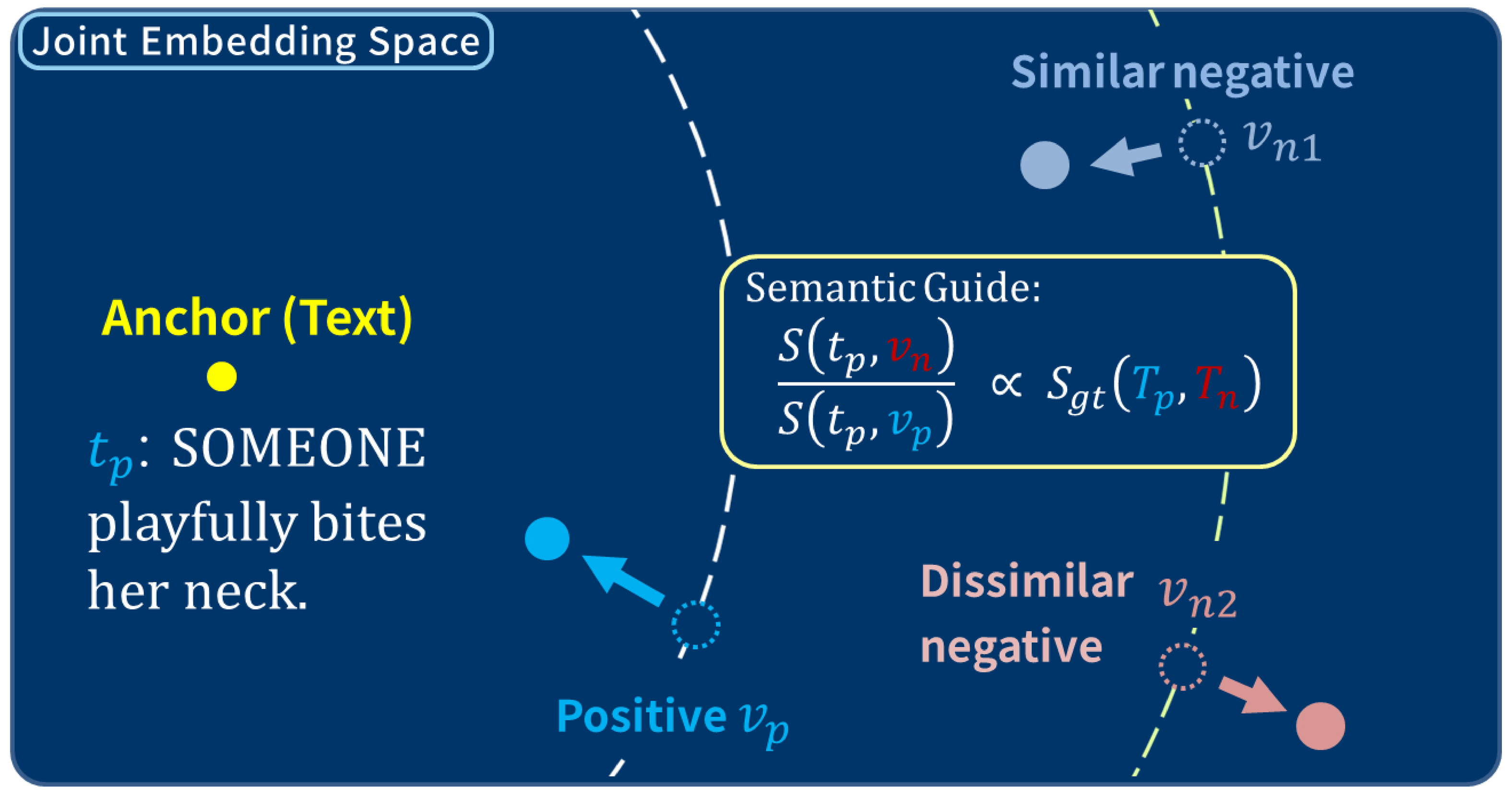


Proposed semantic relevance

$$S_{gt}(T_A, T_B) = 1 - \arccos(E(T_A) \cdot E(T_B))$$

Proposed method

Objectives: Place each negative embeddings according to proposed semantic relevance



Dotted circle: learned embeddings by ranking loss
Solid circle: learned embeddings by the proposed method

Proposed loss for a cross-modal triplet

$$loss_{t \to v}(p, n) = \left\{ \log \frac{S(t_p, v_n)}{S(t_p, v_p)} - \log S_{gt}(T_p, T_n) \right\}^2$$

Overall loss by applying bidirectional loss and hardest negative sampling

$$\mathcal{L}_{cross} = \sum_{p \in D} \left\{ \begin{array}{l} \max_{n \in N_p} loss_{t \to v}(p, n) \\ + \max_{n \in N_p} loss_{v \to t}(p, n) \end{array} \right\}$$

Experimental results

Present a close alignment between the learned metric space and the semantic space

Qualitative results: semantically similar videos are embedded close to each other

Text query : a man is <u>riding a horse</u> at top speed <u>in a race</u>

Comparison text-to-video retrieval result of ranking loss (Rank) and the proposed method on MSR-VTT (Top-5)

Quantitative results: Improve text-to-video retrieval performance on six video-text datasets

Dataset	Loss	Text to video R@1	R@5	R@10	MedR	MeanR	Sum of R	Gain
LSMDC	Rank	10.6	25.5	33.1	33	122.8	69.2	+16.5%
	Semantic	12.7	29.2	38.7	25	103.2	80.6	
DiDeMo	Rank	13.8	33.4	44.8	15	68.2	92.0	+11.6%
	Semantic	15.1	37.2	50.4	10	45.1	102.7	
ANet	Rank	12.8	34.5	48.6	11	68.7	95.9	+9.2%
	Semantic	14.2	37.8	52.8	9	46.6	104.7	
VTT	Rank	8.6	24.4	34.1	29	210.0	67.1	+7.6%
	Semantic	9.1	26.2	37.0	22	163.8	72.2	
TGIF	Rank	5.9	14.8	20.6	122	729.1	41.3	+4.0%
	Semantic	6.0	15.4	21.6	93	563.9	43.0	
VATEX	Rank	35.2	71.5	80.8	2	18.9	187.4	+2.0%
	Semantic	35.3	73.2	82.7	2	15.8	191.2	

Comparison text-to-video retrieval result of ranking loss (Rank) and the proposed method (Semantic)