T2V2T: Text-to-Video-to-Text Fusion for Text-to-Video Retrieval

Jonghee Kim[1] Youngwan Lee[1,2] Jinyoung Moon[1,3]

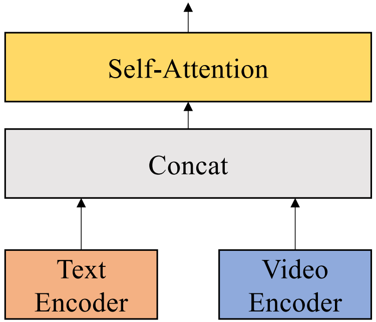
[1]Electronics and Telecommunications Research Institute (ETRI)
[2]Korea Advanced Institute of Science and Technology (KAIST)
[3]University of Science and Technology (UST)

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA
@O-DRUM

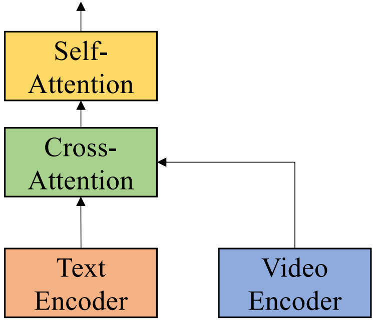
Motivation

Video-Language Transformers based on Joint Encoder

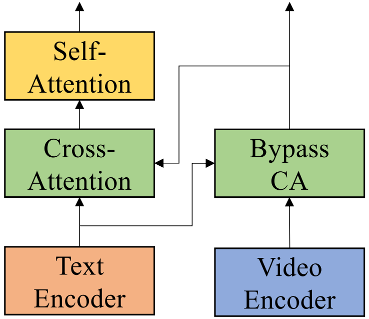
- Self-attention-based fusion
  - A full interaction with high computational cost (L^2)
- Unidirectional fusion
  - A one-way interaction without text-to-video (T2V) interaction
- T2V2T fusion
  - An efficient bidirectional fusion (an alternative to self-attention-based fusion)



Self-attention-based fusion



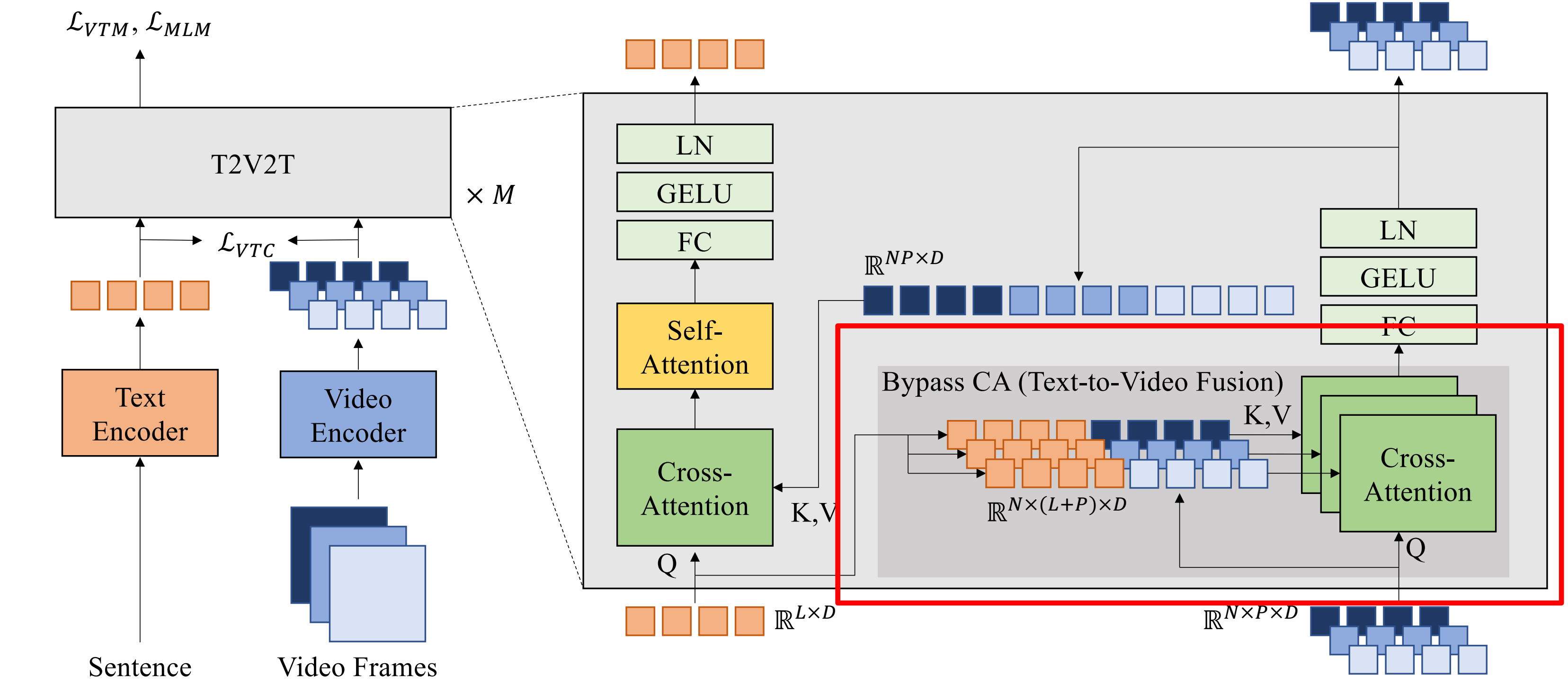
Unidirectional fusion (V2T)



Text-to-Video-to-Text fusion (T2V2T)

Overall Architecture

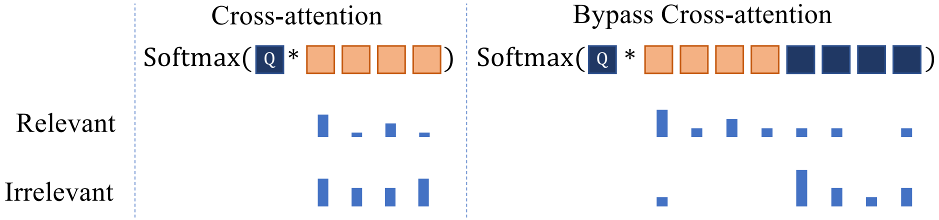
T2V2T Fusion

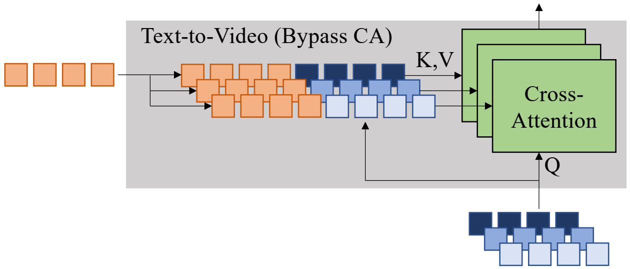


Bypass Cross-attention

Comparison with Cross-Attention

- Cross-attention
  - Associates all frames with the given sentence without regard to the correlation between the given sentence and each frame.
- Bypass cross-attention
  - Considering that only a subset of frames is relevant to the given sentence, we introduce a bypass mechanism in CA
  - It allows frame features to be associated with themselves in the key, instead of text features, if they are irrelevant.





Text-to-Video-to-Text Fusion

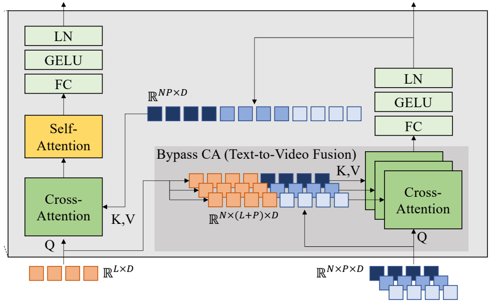
Text-to-Video Fusion followed by Video-to-Text Fusion

- Text-to-Video Fusion
  - Bypass CA incorporates video features into keys and values to prevent modulation by irrelevant text.

$$F_{T2V}[i] = \mathrm{X-Attn}(F_V[i], F_T || F_V[i])$$

- Video-to-Text Fusion: V2T Cross-attention + Self-attention applied to text features

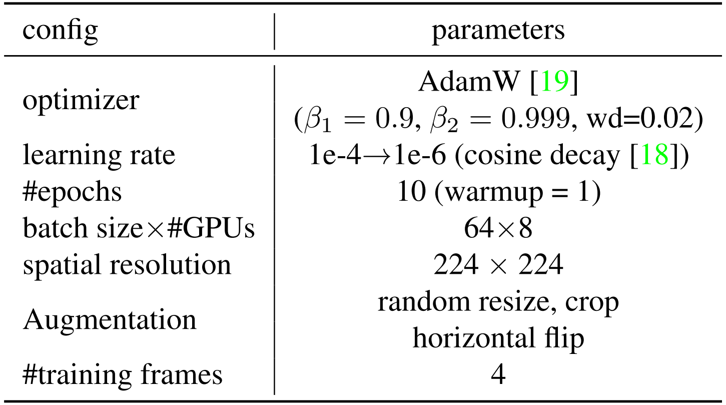
$$F_{T2V2T} = \mathrm{S-Attn}\left(\mathrm{X-Attn}(F_T, F_{T2V})\right)$$


LN
GELU
FC
Self-Attention
Cross-Attention
K,V
Q
$\mathbb{R}^{L\times D}$
$\mathbb{R}^{NP\times D}$
LN
GELU
FC
Bypass CA (Text-to-Video Fusion)
K,V
Cross-Attention
Q
$\mathbb{R}^{N\times(L+P)\times D}$
$\mathbb{R}^{N\times P\times D}$

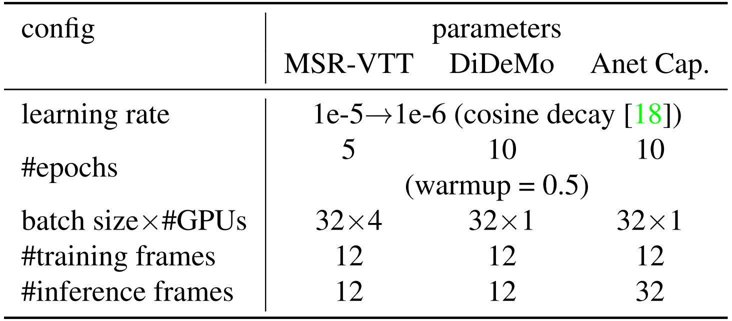
Experimental Results

Experiment setup

- Video-language pre-training
  - CC3M, WebVid-2.5M
- Text-to-video retrieval
  - MSR-VTT, DiDeMo, ActivityNet Captions



config	parameters
optimizer	AdamW [19] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, wd=0.02)
learning rate	1e-4→1e-6 (cosine decay [18])
#epochs	10 (warmup = 1)
batch size×#GPUs	64×8
spatial resolution	224 × 224
Augmentation	random resize, crop horizontal flip
#training frames	4



config	parameters		
	MSR-VTT	DiDeMo	Anet Cap.
learning rate	1e-5→1e-6 (cosine decay [18])		
#epochs	5	10	10
	(warmup = 0.5)		
batch size×#GPUs	32×4	32×1	32×1
#training frames	12	12	12
#inference frames	12	12	32

Experimental Results

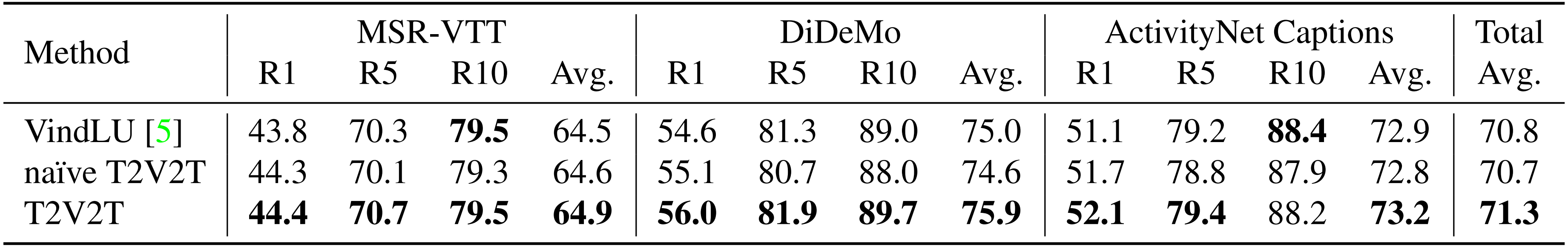
Text-to-video retrieval

Method	#PT	MSRVTT				DiDeMo				ActivityNet Captions			
		R1	R5	R10	Avg.	R1	R5	R10	Avg.	R1	R5	R10	Avg.
ClipBERT [13]	5.6M	22.0	46.8	59.9	42.9	20.4	48.0	60.8	43.1	21.3	49.0	63.5	44.6
Frozen [2]	5.5M	31.0	59.5	70.5	53.7	31.0	59.8	72.4	54.4	-	-	-	-
ALPRO [15]	5.5M	33.9	60.7	73.2	55.9	35.9	67.5	78.8	60.7	-	-	-	-
BridgeFormer [9]	5.5M	37.6	64.8	75.1	59.2	37.0	62.2	73.9	57.7	-	-	-	-
Singularity [12]	5.5M	39.9	67.3	76.0	61.1	49.2	77.5	85.4	70.7	45.9	73.3	83.8	67.7
VindLU [5]	5.5M	43.8	70.3	**79.5**	64.5	54.6	81.3	89.0	75.0	51.1	79.2	**88.4**	72.9
T2V2T (Ours)	5.5M	**44.4**	**70.7**	**79.5**	**64.9**	**56.0**	**81.9**	**89.7**	**75.9**	**52.1**	**79.4**	88.2	**73.2**
MMT [8]	136M	25.8	57.2	69.3	50.8	-	-	-	-	28.7	61.4	94.5	61.5
TACo [29]	120M	28.4	57.8	71.2	52.5	-	-	-	-	30.4	61.2	93.4	61.7
SupportSet [22]	120M	30.1	58.5	69.3	52.6	-	-	-	-	29.2	61.6	94.7	61.8
Singularity [12]	17M	42.7	69.5	78.1	63.4	53.1	79.9	88.1	73.7	48.9	77.0	86.3	70.7
VindLU [5]	17M	45.3	69.9	79.6	64.9	59.2	84.1	89.5	77.6	54.4	80.7	89.0	74.7
CLIP4Clip [21]	400M	44.5	71.4	81.6	65.8	43.4	70.2	80.6	64.7	40.5	72.4	98.2	70.4
VindLU [5]	25M	46.5	71.5	80.4	66.1	61.2	85.8	91.0	79.3	55.0	81.4	89.7	75.4
OmniVL [26]	17M	47.8	74.2	83.8	68.6	52.4	79.5	85.4	72.4	-	-	-	-

Experimental Results

Ablation study to evaluate the effectiveness of Bypass CA

- Baseline: VindLU
- Naïve T2V2T: VindLU + V2T cross-attention
- T2V2T: VindLU + Bypass cross-attention



Method	MSR-VTT				DiDeMo				ActivityNet Captions				Total
	R1	R5	R10	Avg.	R1	R5	R10	Avg.	R1	R5	R10	Avg.	Avg.
VindLU [5]	43.8	70.3	**79.5**	64.5	54.6	81.3	89.0	75.0	51.1	79.2	**88.4**	72.9	70.8
naïve T2V2T	44.3	70.1	79.3	64.6	55.1	80.7	88.0	74.6	51.7	78.8	87.9	72.8	70.7
T2V2T	**44.4**	**70.7**	**79.5**	**64.9**	**56.0**	**81.9**	**89.7**	**75.9**	**52.1**	**79.4**	88.2	**73.2**	**71.3**

Conclusion

- We conducted an investigation into the text-to-video (T2V) interaction, which suffers from an imbalance between the number of video and text embeddings.
  (32 vs 784)

- We have proposed a novel fusion method, termed T2V2T fusion,
  which incorporates Bypass CA.

- T2V2T has achieved SOTA text-to-video retrieval results on MSR-VTT, DiDeMo, and ActivityNet Captions.