Lightweight Attentional Feature Fusion: A New Baseline for Text-to-Video Retrieval

Fan Hu , Aozhu Chen , Ziyue Wang, Fangming Zhou, Jianfeng Dong, Xirong Li
MoE Key Lab of DEKE, Renmin University of China
AIMC Lab, School of Information, Renmin University of China
College of Computer and Information Engineering, Zhejiang Gongshang University

EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022
October 23-27, 2022, Tel Aviv

1. Summary

Background:

Given video/text samples represented by diverse features, we need an **optimal** way to combine these features.

Challenges:

1. Previous feature fusion SOTAs are either **too simple** (average, concatenation) or **too heavy** (Transformers).
2. Previous researches are **modality specific** and lack interpretation.

Our Solution:

Lightweight Attentional Feature Fusion (**LAFF**)

2. Proposed LAFF

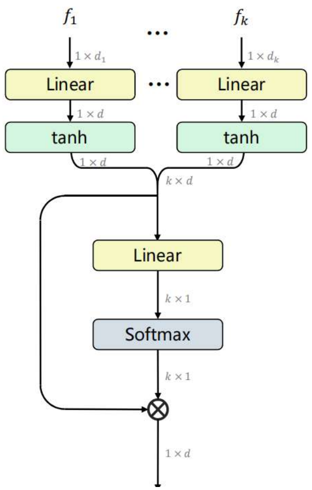


Fig.1. LAFF block

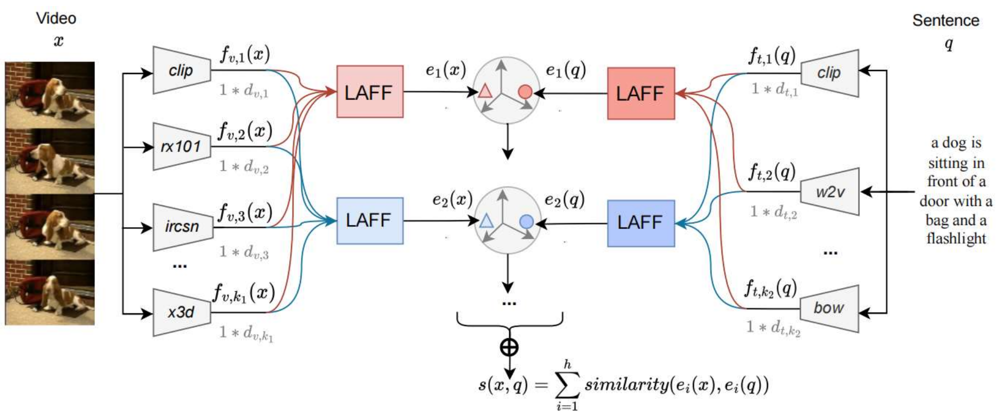


Fig.2. Paired LAFFs

2.1 The LAFF Block

- Feature transformation layer to rectify the diverse features to the same length.
- Linear and Softmax to obtain the weights of different features.

2.2 Paired LAFFs for Text-to-Video Retrieval

- LAFF for both text and video feature fusion.
- Multi-head idea to get the fusion feature for different paired LAFFs.

3. Experiments

3.1 Verification Study on MSRVTT

Feature Fusion block	Parameters	FLOPs (M)
MHSA	$D \times d + 4 \times d^2$	94.90
LAFF	$D \times d + d$	27.80

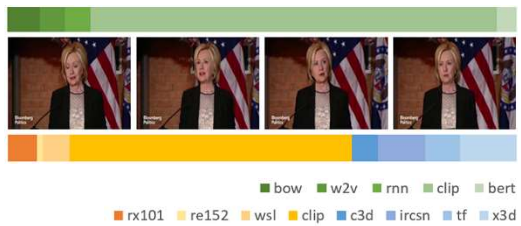
D: the overall dimension of the input features
d: the dimension of the output features

LAFF is **Lightweight**

Fusion block	R1	R5	R10	Medr	mAP
MHSA	18.8	43.0	54.6	8	0.305
LAFF	**23.7**	**49.1**	**60.6**	**6**	**0.358** (15.5%↑)

LAFF is **effective**

Hillary Clinton gives a speech on race



LAFF can **select features** according to the attentional weights

3.2 Comparison with SOTA

Model	MV-test3k			MV-test1k			MSVD			TGIF			VATEX		
	R1	*R5*	*R10*	*R1*	*R5*	*R10*	*R1*	*R5*	*R10*	*R1*	*R5*	*R10*	*R1*	*R5*	*R10*
CLIP-FT (*this paper*)	27.7	53.0	64.2	39.7	67.8	78.4	44.6	74.7	84.1	21.5	40.6	49.9	53.3	87.5	94.0
The same video and text feature as ours															
JE [41] (uniform weights)	21.2	46.5	58.4	36.0	65.9	76.4	35.9	71.0	81.8	18.7	37.5	47.1	50.2	88.7	95.4
JE (0.8 for *clip-ft*)	26.1	51.7	63.3	41.2	73.2	82.5	39.4	69.9	79.4	21.7	41.3	50.9	54.1	89.0	95.0
JE (0.9 for *clip-ft*)	25.9	51.4	63.0	40.9	72.7	82.1	38.8	69.7	78.9	21.3	40.9	50.3	53.5	88.3	94.6
W2VV++ [28]	23.0	49.0	60.7	39.4	68.1	78.1	37.8	71.0	81.6	22.0	42.8	52.7	55.8	91.2	96.0
SEA [31]	19.9	44.3	56.5	37.2	67.1	78.3	34.5	68.8	80.5	16.4	33.6	42.5	52.4	90.2	95.9
MMT [19]	24.9	50.5	62.0	39.5	68.3	78.3	40.6	72.0	81.7	22.1	42.2	51.7	54.4	89.2	95.0
LAFF	28.0	53.8	64.9	42.2	70.7	**81.2**	45.2	75.8	84.3	24.1	44.7	54.3	57.7	91.3	95.9
LAFF-ml	**29.1**	**54.9**	**65.8**	**42.6**	**71.8**	81.0	**45.4**	**76.0**	**84.6**	**24.5**	**45.0**	**54.5**	**59.1**	**91.7**	**96.3**
Comparison with arXiv SOTA															
CLIP2Video [17]	n.a	n.a	n.a	44.5	71.3	80.6	44.7	74.8	83.7	n.a	n.a	n.a	54.8	89.1	95.1
LAFF	n.a	n.a	n.a	**45.8**	**71.5**	**82.0**	**45.4**	**75.5**	**84.1**	n.a	n.a	n.a	**58.3**	**91.7**	**96.3**

- With the same video and text feature, LAFF perform best compare to Baselines.
- Including the global video/text features extracted by CLIP2Video (arXiv SOTA), LAFF can flexibly harness new and more powerful features.