Learning Linguistic Association Towards Efficient Text-Video Retrieval

Sheng Fang[1,2], Shuhui Wang[1,3*], Junbao Zhuo[1], Xinzhe Han[2,1] and Qingming Huang[2,1,3]
[1]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS,
[2]University of Chinese Academy of Sciences, [3]Peng Cheng Laboratory

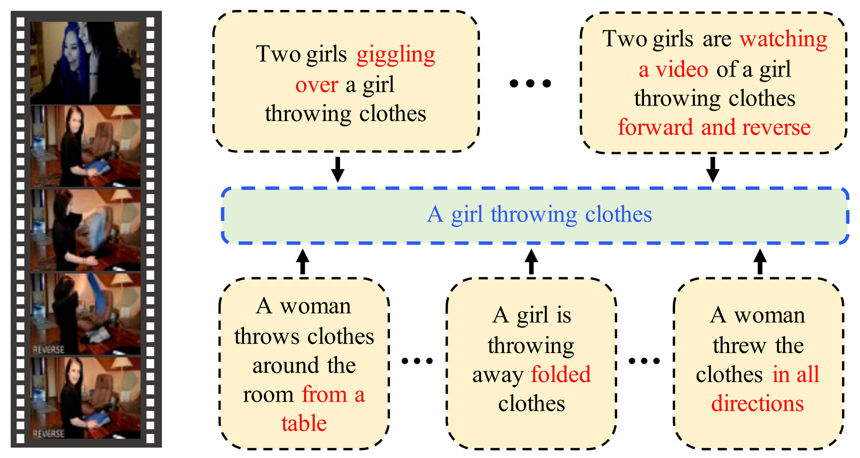
EUROPEAN CONFERENCE ON COMPUTER VISION
ECCV
TEL AVIV 2022
October 23-27, 2022

Imbalance Problem in TVR

➢ Video delivers richer content than text

- **Nature Side**
  - Video → a consecutive photometric record of events in a physical world.
  - Text → the abstract description of the events that a person sees or experiences.
- **Language Side**
  - Video → describe an event without missing details of actions, attributes and objects.
  - Text → depicted by different individuals and may have different focuses and language habits.

➢ Imbalance makes it difficult to align text and video



Contributions

- We propose a general framework, LINguistic ASsociation (LINAS), which encourages the model to learn the ability of semantics enrichment for better aligning two modalities.
- We introduce Knowledge Distillation and further propose Adaptive Distillation strategy for suppressing the spurious correlation in the teacher model.
- Extensive experiments demonstrate the effectiveness, efficiency, and generalization ability of LINAS. Our code is now available.

➢ Code Release

Experiments

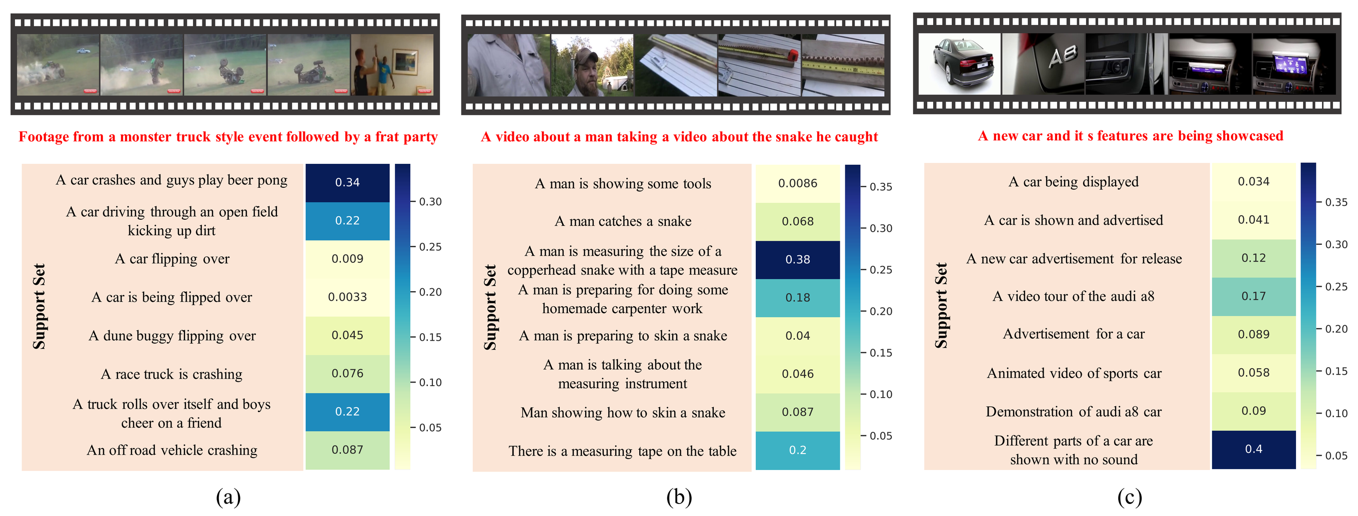
➢ Comparison

Experiments on benchmark dataset MSR-VTT.

Method	Text2Video					Video2Text					SumR
	R@1	R@5	R@10	MedR	mAP	R@1	R@5	R@10	MedR	mAP	
VSE++[11]	5.7	17.1	24.8	65	-	10.2	25.4	35.1	25	-	118.3
Mithun *et al.*[29]	7.0	20.9	29.7	38	-	12.5	32.1	42.4	16	-	144.6
W2VV[8]	6.1	18.7	27.5	45	-	11.8	28.9	39.1	21	-	132.1
CE[23]	10.0	29.0	41.2	16	-	15.6	40.9	55.2	8.3	-	191.9
HGR[5]	9.2	26.2	36.5	24	-	15.0	36.7	48.8	11	-	172.4
Dual Encoding[9]	11.0	29.3	39.9	19	20.3	19.7	43.6	55.6	8	9.3	199.0
LINAS - Dual Encoding	11.9	31.0	42.1	17	21.6	22.0	46.9	59.2	6	10.4	213.1
Hybrid Space[10]	11.6	30.3	41.3	17	21.2	22.5	47.1	58.9	7	10.5	211.7
LINAS - Hybrid Space	12.3	31.6	42.8	16	22.1	22.3	47.8	60.4	6	10.6	217.2
CE+[7]	14.4	37.4	50.2	**10**	-	22.7	52.6	66.3	5	-	243.6
TeachText - CE+[7]	14.9	38.3	51.5	**10**	-	**24.9**	54.1	67.6	5	-	251.3
LINAS - CE+	**15.2**	**38.9**	**52.0**	**10**	-	24.7	**55.2**	**68.0**	**4**	-	**254.0**

➢ Visualization

Text-video pairs and the attention weights of support set captions.



➢ Ablation Study

Experiments about distillation strategy.

	Distillation Loss				Text2Video					Video2Text					SumR
	$\mathcal{L}_{D_{text}}$	$\mathcal{L}_{D_{video}}$	$\mathcal{L}_{D_{rel}}$	$\mathcal{L}'_{D_{rel}}$	R@1	R@5	R@10	MedR	mAP	R@1	R@5	R@10	MedR	mAP	
1					10.9	29.3	39.8	20	20.2	19.5	42.8	55.8	8	9.3	199.0
2	✓				11.3	30.0	40.8	18	20.8	21.1	44.6	56.7	7	10.1	204.4
3			✓		11.3	30.1	41.1	18	20.9	20.8	44.5	58.2	7	9.8	205.9
4	✓	✓			11.7	30.6	41.6	**17**	21.3	21.9	45.2	58.3	7	10.2	209.3
5	✓		✓		11.5	30.2	41.1	18	21.0	20.4	45.8	57.7	7	10.2	206.7
6	✓	✓		✓	11.5	30.2	41.1	18	21.0	21.8	45.5	58.2	7	10.0	208.3
7	✓	✓	✓		**11.9**	**31.0**	**42.1**	**17**	**21.6**	**22.0**	**46.9**	**59.2**	**6**	**10.4**	**213.1**
Teacher Model					19.0	44.6	57.7	7	31.3	23.7	39.0	46.9	13	18.9	231.0

➢ General Applicability

Experiments on different pretraining baseline methods.

Method	Dataset	Text2Video				Video2Text				SumR
		R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
ClipBERT[17]	MSR-VTT 1k-A	22.0	46.8	59.9	6	-	-	-	-	-
MMT[13]		25.8	57.3	69.3	4	26.1	57.8	68.5	4	304.8
LINAS - MMT		27.1	59.8	71.7	4	28.3	60.3	72.0	3	319.2
Frozen in time[2]	MSVD	33.7	64.7	76.3	3	-	-	-	-	-
CLIP4Clip[24]		46.2	76.1	84.6	2	-	-	-	-	-
LINAS - CLIP4Clip		46.7	76.8	85.6	2	47.3	75.0	83.2	2	414.6

Comparable results with 3× speed-up efficiency.

Teacher Model	Student Model	Text2Video				Video2Text				SumR
		R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
-	Base	9.7	26.8	37.0	23	17.7	40.0	51.7	10	182.9
Base	Base	10.3	27.9	38.9	**19**	19.3	42.9	54.1	**8**	193.4
Dual Encoding[9]	Base	10.7	28.9	**39.9**	**19**	**20.0**	**43.9**	**56.4**	**8**	**199.8**
-	Dual Encoding[9]	**10.9**	**29.3**	39.8	20	19.5	42.8	55.8	**8**	199.0

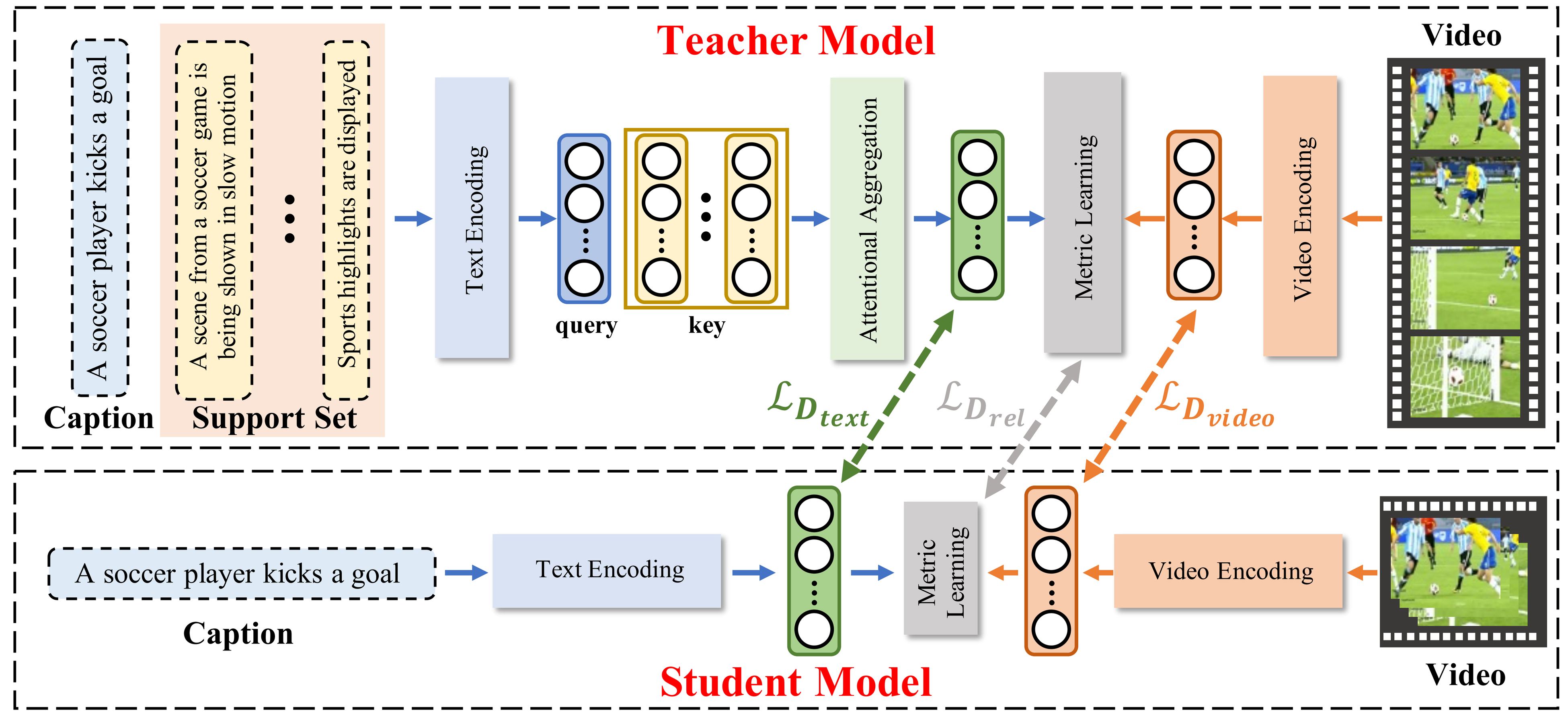
Method

➢ Teacher Model

- **Support Set Construction.** (1) Descriptions belonging to the same video with the query caption; (2) Captions of semantically similar videos.
- **Attentional Aggregation Module.** Combine the complementary semantics to the query caption.

$$x_i^t = q_i + \sum_{n=1}^{N} \frac{\exp(Q(q_i)^T K(k_i^n))}{\sum_{l=1}^{N} \exp(Q(q_i)^T K(k_i^l))} k_i^n,$$

- **Training.** Requires additional support captions as input.



➢ Student Model

- **Training with Knowledge Distillation.** (1) Feature-level distillation; (2) Relational distillation.

$$\mathcal{L}_{D_{text}} = \sum_{i=1}^{B}(\|x_i^t - x_i^s\|_2^2), \quad \mathcal{L}_{D_{video}} = \sum_{i=1}^{B}(\|y_i^t - y_i^s\|_2^2), \quad \mathcal{L}_{D_{rel}} = \sum_{i=1}^{B}\sum_{j=1}^{B} m(i,j) L_\delta(S^t(i,j), S^s(i,j)),$$

- **Inference.** Brings no extra computation cost.

➢ Adaptive Distillation

- **Spurious Correlation.** Not all the correlations from the teacher model are reliable.
- **Algorithm.** Adopt EM (Expectation-Maximization) Algorithm to iteratively optimize m and θ.

Algorithm 1: Adaptive Distillation

Create a mask m which is uniformly initialized. θ represents the model parameters.
while *not converged* **do**
  $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{train}(\theta, m)$;
  $m \leftarrow m - \eta_m \nabla_m \mathcal{L}_{val}(\theta, m)$;
end
Based on the learned mask m, retrain the model parameters θ from scratch.

$$\mathcal{L}_{val}(\theta, m) = \sum_{i=1}^{B}\sum_{j=1}^{B} \frac{1}{S^t(i,j)} m(i,j) L_\delta(S^t(i,j), S^s(i,j)).$$

- **Learning Process.** Model tends to transfer the relational knowledge of the diagonal elements from the teacher model.