Hybrid-Learning Video Moment Retrieval across Multi-Domain Labels

Weitong Cai[1] Jiabo Huang[1] Shaogang Gong[1]

[1]Queen Mary University of London

BMVC 2022

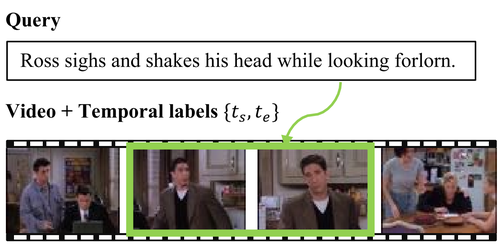
Background

Video Moment Retrieval (VMR)

- A task to locate a temporal moment (start & end points) in a long and untrimmed *video* according to a natural language *query* sentence.

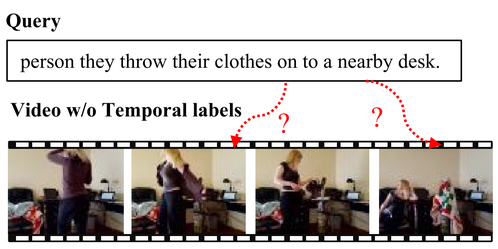
Existing methods (two approaches):

1) Fully-Supervised VMR



- **Have** temporal labels in the training
- Harder to annotate
- Ambiguity-prone and sensitive to subjective bias
- Poor generalisation

2) Weakly-Supervised VMR



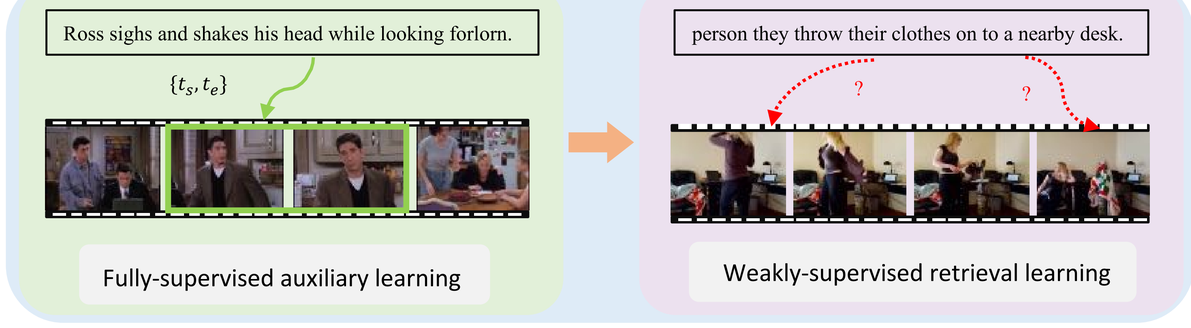
- **No** temporal labels in the training
- Harder to derive semantically plausible video-text correspondences
- Overpass existing available temporal boundary labels; lead to waste

Ideas and Contributions

Hybrid-Learning Video Moment Retrieval

❖ **Overall Idea**

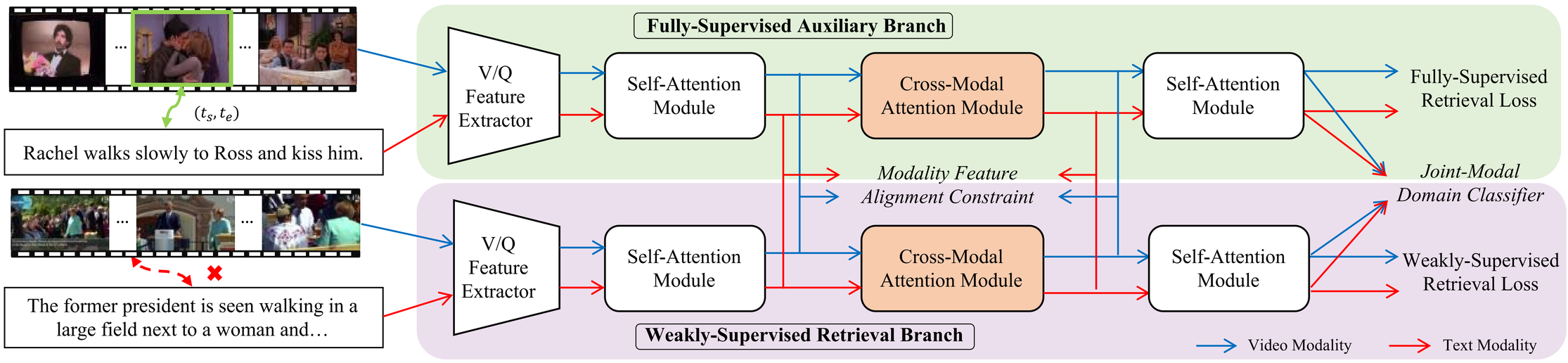
- Optimise weakly-supervised retrieval learning of visual-textual correlations in a target domain by sharing knowledge on video-text alignment learned from fully-supervised auxiliary learning in a source domain.



❖ **Main Contributions**

a) **A new approach** to VMR: Hybrid-Learning VMR
b) **A multiplE branch Video-text Alignment model (EVA)** to transfer temporal label information as knowledge across domains/tasks
c) **Competitive results** against the state-of-the-art methods

Method



Multi-Branch Hybrid-Learning

❖ **Weakly-Supervised Retrieval Branch**

1) Within-modal self-attention module

$$\mathcal{R}(Y,X) = \text{softmax}(YW^{q\top}W^{k}X^{\top}/\sqrt{d}), \quad Att(Y,X) = \text{FC}(Y + \mathcal{R}(Y,X)XW^{v\top}),$$

$$V \leftarrow Att^{V}(V,V), \quad Q \leftarrow Att^{Q}(Q,Q),$$

2) Weakly-supervised retrieval loss

$$\mathcal{L}_w = 2 * (-\log P_m(V|Q)) - \log(1 - P_m(V|Q^{-})) - \log(1 - P_m(V^{-}|Q)).$$

❖ **Fully-Supervised Retrieval Branch**

1) Cross-modal attention module (sharing parameters)

$$V \leftarrow Att^{Q\rightarrow V}(V,Q), \quad Q \leftarrow Att^{V\rightarrow Q}(Q,V),$$

2) Fully-supervised retrieval loss

$$\mathcal{L}_f = \lambda_r \mathcal{L}_r^f + \mathcal{L}_{bce}^f, \quad \mathcal{L}_r^f = H(P_s, I_s^f) + H(P_e, I_e^f),$$

Multi-Modal Feature Alignment across Tasks and Domains

❖ **Modality Feature Alignment Constraint**

$$M(D^s, D^t)^2 = \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s} K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} K(\boldsymbol{y}_i, \boldsymbol{y}_j) - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} K(\boldsymbol{x}_i, \boldsymbol{y}_j),$$

$$\mathcal{L}_{align} = (\lambda_{vid} M(V_b^f, V_b^w)^2 + M(Q_b^f, Q_b^w)^2) + (\lambda_{vid} M(V_a^f, V_a^w)^2 + M(Q_a^f, Q_a^w)^2),$$

❖ **Joint-Modal Domain Classifier**

$$G_d(J) = \text{softmax}(\text{FC}_1(\text{FC}_2(J))).$$

$$\mathcal{L}_{domain} = -\log(1 - G_d(J^f)) - \log G_d(J^w),$$

Overall loss

$$\mathcal{L} = \mathcal{L}_w + \lambda_f \mathcal{L}_f + \lambda_{align}\mathcal{L}_{align} - \lambda_{domain}\mathcal{L}_{domain},$$

Experiments

➢ Comparisons with fully-supervised methods

Method	Source	Target	Charades			Anet		
			IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.1	IoU=0.3	IoU=0.5
2D-TAN [48]	✓	✗	14.65	4.30	1.26	40.16	28.71	17.29
XML [21]	✓	✗	32.49	18.27	8.87	31.78	17.18	9.27
MMN [39]	✓	✗	11.45	3.06	0.86	42.19	24.57	13.09
EVA	✓	✓	**62.01**	**40.21**	**18.22**	**74.09**	**49.89**	**29.43**

➢ Component ablation study

Method	mIoU	IoU=0.1	IoU=0.3	IoU=0.5
WR	32.96	71.48	48.06	28.74
WR + fine-tune	33.14	72.59	48.86	28.21
WR + unlabelled source	33.01	72.18	48.18	28.01
WR + FA	33.11	71.62	48.46	28.56
WR + FA + Align	33.83	73.35	49.60	28.80
WR + FA + Domain	33.66	73.14	49.32	28.92
WR + FA + Align + Domain	**34.27**	**74.09**	**49.89**	**29.43**

➢ Comparisons with weakly-supervised methods

Method	Source	Target	IoU=0.3	IoU=0.5	IoU=0.7
TGA [32]	✗	✓	29.68	17.04	6.93
SCN [22]	✗	✓	42.96	23.58	9.97
LoGAN [37]	✗	✓	51.67	34.68	14.54
BAR [40]	✗	✓	44.97	27.04	12.23
RTBPN [49]	✗	✓	60.04	32.36	13.24
VLANet [31]	✗	✓	45.24	31.83	14.17
CCL [50]	✗	✓	-	33.21	15.68
CRM [18]	✗	✓	53.66	34.76	16.37
EVA	✓	✓	**62.01**	**40.21**	**18.22**

(a) Evaluated on Charades

Method	Source	Target	Split	IoU=0.1	IoU=0.3	IoU=0.5
WS-DEC [13]	✗	✓	val_1	62.71	41.98	23.34
WSLLN [16]	✗	✓	val_1	75.4	42.8	22.7
BAR [40]	✗	✓	val_1	-	49.03	30.73
CRM [18]	✗	✓	val_1	**76.66**	**51.17**	**31.67**
EVA	✓	✓	val_1	70.79	46.23	28.00
SCN [22]	✗	✓	val_2	71.48	47.23	29.22
RTBPN [49]	✗	✓	val_2	73.73	49.77	29.63
CCL [50]	✗	✓	val_2	-	50.12	31.07
CRM [18]	✗	✓	val_2	**81.61**	**55.26**	**32.19**
EVA	✓	✓	val_2	74.09	49.89	29.43

(b) Evaluated on Anet

➢ Comparisons on OOD splits

Dataset	Method	Source	Target	IoU=0.3	IoU=0.5	IoU=0.7
Anet	WS-DEC [13]	✗	✓	17.00	7.17	1.82
	CRM [18]	✗	✓	22.77	10.31	-
	EVA	✓	✓	**23.11**	**11.29**	**4.32**
Charades	WS-DEC [13]	✗	✓	35.86	23.67	8.27
	EVA	✓	✓	**47.83**	**31.71**	**12.76**