腾讯优图

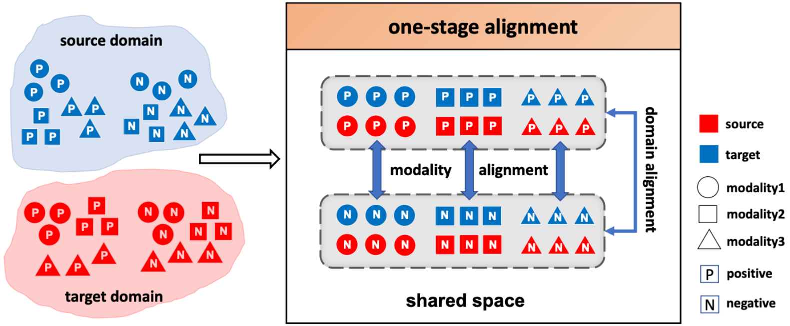
OSAN: A One-Stage Alignment Network to Unify Multimodal Alignment and Unsupervised Domain Adaptation

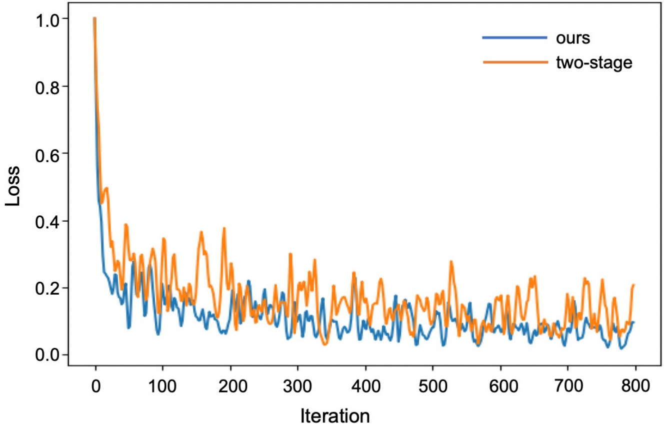
Ye Liu, Lingfeng Qiao, Changchong Lu, Di Yin, Chen Lin, Haoyuan Peng, Bo Ren
Tencent Youtu Lab

paper tag: TUE-AM-338
poster ID: 338

Motivation

腾讯优图





Key issues

- how to align the source and target domains and remit domain discrepancy.
- how to align multiple modalities and leverage multimodal information.

Traditional method

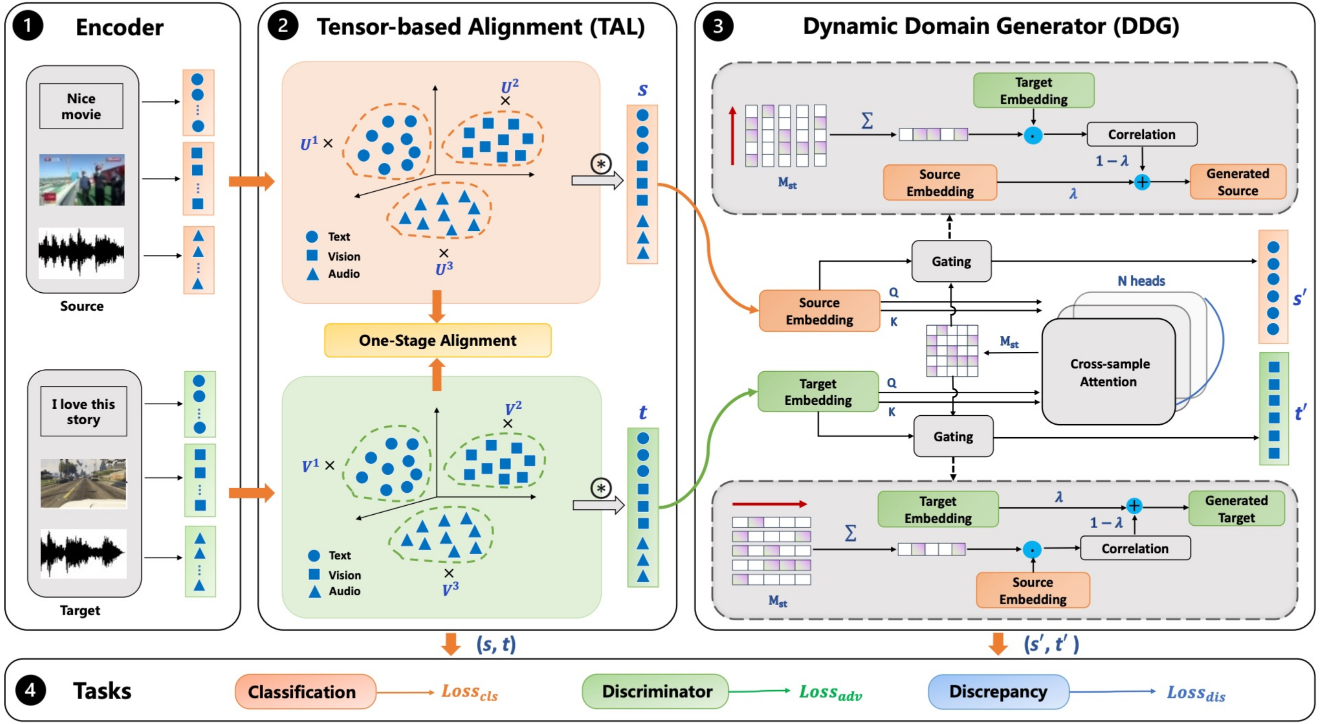
- Most existing works address these two problems in two consecutive stages: multimodal alignment followed by domain adaptation, or vice versa.

weakness

inability to preserve the relations between modalities while performing domain adaptation

OSAN: Framework

腾讯优图



The schematic diagram of our OSAN algorithm

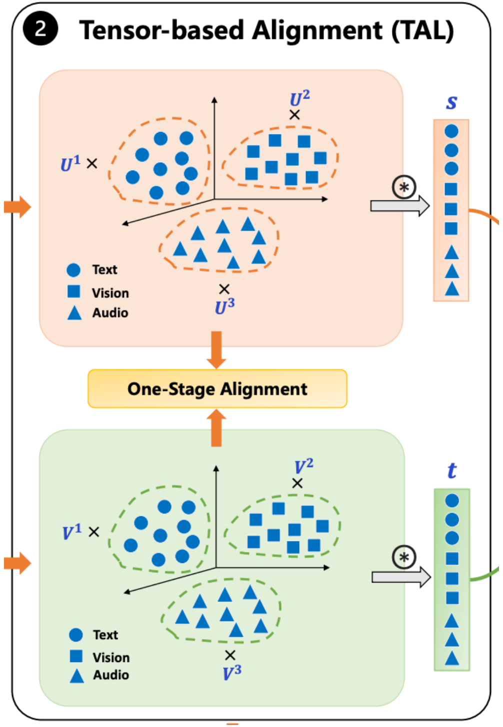
OSAN: Tensor-based Alignment

腾讯优图

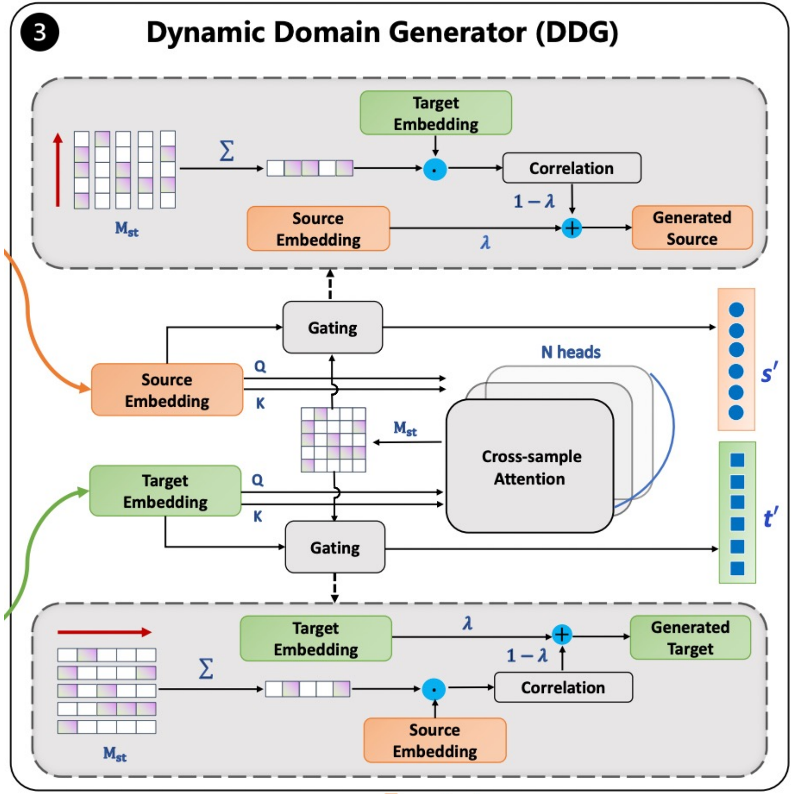
$$
\begin{aligned}
&\left\{\mathbf{U}^{(n)}\big|_{n=1}^{N}, \mathbf{V}^{(n)}\big|_{n=1}^{N}\right\} \\
&= \max[\![(\underline{\mathbf{X}} \prod_{n=1}^{N} \times_n \mathbf{U}^{(n)^T}) \otimes (\underline{\mathbf{Y}} \prod_{n=1}^{N} \times_n \mathbf{V}^{(n)^T});]\!] \\
&s.t.\ (\underline{\mathbf{X}} \prod_{n=1}^{N} \times_n \mathbf{U}^{(n)^T})^T_{(N+1)} (\underline{\mathbf{X}} \prod_{n=1}^{N} \times_n \mathbf{U}^{(n)^T})_{(N+1)} = \mathbf{I} \\
&\quad (\underline{\mathbf{Y}} \prod_{n=1}^{N} \times_n \mathbf{V}^{(n)^T})^T_{(N+1)} (\underline{\mathbf{Y}} \prod_{n=1}^{N} \times_n \mathbf{V}^{(n)^T})_{(N+1)} = \mathbf{I}
\end{aligned}
\tag{1}
$$

Motivation: perform multimodal alignment and domain adaptation at the same time.

Objectiveness : To perform multimodal alignment, TAL aims to find pairs of linear transformations for each modality of source and target domains to project samples of two sets into low dimensional subspaces. During this process, we establish an interaction between domain and modality by maximizing a statistical measurement of covariance given a normalized standard deviation



OSAN: Dynamic Domain Generator



How: DDG explicitly captures commonality and abandons the specialty of domains. By highlighting this commonality, we make the domain discriminator focus on commonality rather than full information, which helps our model learn a domain invariant common representation space.

Significance: Samples from new domains and raw source and target domains are fed to domain discriminator, by which the domain discriminator is guided by the hard label information and well-designed soft domains. Each sample from these soft domains explores the intrinsic structure of data distribution from raw domains and enriches feature patterns by the interaction of two domains.

Experiment: Multimodal Sentiment Analysis

腾讯优图

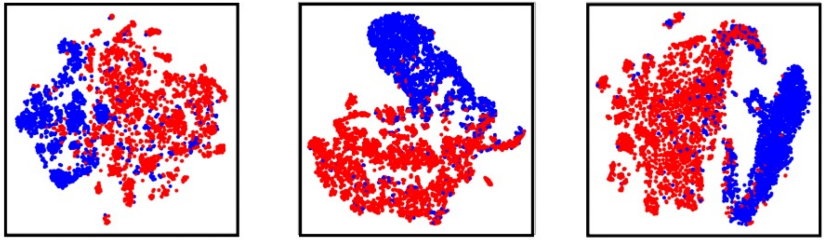
Table 1. Multimodal sentiment analysis results on CMU-MOSI and CMU-MOSEI. †: results come from [7]; ‡: results come from [36]; ◊: results come from [6]; ↓: the lower the better.

Methods	CMU-MOSEI ⟶ CMU-MOSI					CMU-MOSI ⟶ CMU-MOSEI				
	MAE ↓	Corr	Acc-7	Acc-2	F1	MAE ↓	Corr	Acc-7	Acc-2	F1
Direct Transfer	0.794	0.764	39.5	79.7/81.5	79.5/81.4	0.621	0.685	51.3	79.54/82.14	80.84/81.33
Supervised										
TFN [37] †	0.901	0.698	34.9	-/80.8	-/80.7	0.593	0.700	50.2	-/82.5	-/82.1
ICCN [22] †	0.862	0.714	39.0	-/83.0	-/83.0	0.565	0.713	51.6	-/84.2	-/84.2
MISA [7] ‡	0.804	0.764	-	80.79/82.10	80.77/82.03	0.568	0.724	-	82.59/84.23	82.67/83.97
MAG-BERT [20] ‡	0.731	0.789	-	82.50/84.30	82.60/84.30	0.539	0.753	-	83.80/85.20	83.70/85.10
Self-MM [36] ‡	0.713	0.798	-	84.00/85.98	84.42/85.95	0.530	0.765	-	82.81/85.17	82.53/85.30
MMIM [6] ◊	0.700	0.800	46.65	84.14/86.06	84.00/85.98	0.526	0.772	54.24	82.24/85.97	82.66/85.94
UDA										
DAN [12]	0.777	0.774	39.79	80.03/81.71	79.74/81.49	0.614	0.693	51.6	80.24/81.32	81.36/82.47
ADDA [27]	0.784	0.773	40.14	80.12/82.26	80.13/82.32	0.636	0.707	51.4	80.47/81.59	81.53/82.76
MM-SADA [15]	0.787	0.769	40.52	80.9/82.77	80.68/82.63	0.667	0.684	52.1	80.32/81.44	81.26/81.95
MDMN [44]	0.778	0.774	39.65	81.92/82.01	81.97/82.11	0.602	0.712	52.8	82.24/82.38	82.95/83.26
OSAN(TAL + Mixup)	0.753	0.782	42.64	82.44/83.32	82.14/83.21	0.542	0.757	53.14	82.76/82.88	83.13/83.96
OSAN(TAL + DDG)	**0.713**	**0.801**	**46.38**	**83.12/84.58**	**83.02/84.51**	**0.532**	**0.768**	**53.84**	**83.41/84.36**	**83.31/84.47**

Experiment: Video text classification

Table 6. Video text classification results on Text-show.

Methods	Text-news ⟶ Text-show		
	Precision	Recall	F1
Direct Transfer	80.2	77.98	79.08
UDA			
DAN [12]	87.44	80.58	83.87
ADDA [27]	91.66	83.66	87.48
MM-SADA [15]	94.07	83.49	88.46
MDMN [44]	93.54	83.69	88.34
OSAN(TAL + Mixup)	94.42	84.79	89.35
OSAN(TAL + DDG)	**95.03**	**86.44**	**90.53**



(a) Directly transfer (b) Our method w/o TAL&DDG (c) Our method

Figure 5. Visualization of the feature distribution of the target domain for video text classification.

Conclusion

腾讯优图

- To capture the relationship between domain and modality, we propose a one-stage alignment network, called OSAN, to associate domain and modality. In this way, a joint domain-invariant and cross-modal representation space is learned in one stage
- We design a TAL module to bring sufficient interactions between domains and modalities and guide them to utilize complementary information for each other.
- To effectively bridge distinct domains, a DDG module is developed to dynamically construct multiple new domains by combining knowledge of source and target domains and exploring intrinsic structure of data distribution.
- Extensive experiments on two totally different tasks demonstrate the effectiveness of our method compared to the supervised and strongly UDA methods..

Thanks！Q&A