WSDM

MAF: A General Matching and Alignment Framework for Multimodal Named Entity Recognition

Bo Xu
xubo@dhu.edu.cn
School of Computer Science and Technology, Donghua University
Shanghai, China

Shizhou Huang
2202408@mail.dhu.edu.cn
School of Computer Science and Technology, Donghua University
Shanghai, China

Chaofeng Sha*
cfsha@fudan.edu.cn
School of Computer Science, Fudan University
Shanghai Key Laboratory of Intelligence Processing
Shanghai, China

Hongya Wang
hywang@dhu.edu.cn
School of Computer Science and Technology, Donghua University
Shanghai, China

The source code of this paper can be found in https://github.com/xubodhu/MAF

孙凯文 | 导师：杨武 黄丹

Multimodal Named Entity Recognition

It can improve text-based named entity recognition (NER) by using images as additional input. When text information is insufficient, image information can help identify ambiguous named entities.

举例 EXAMPLE

text

Handsome **Rob** after a fish dinner

image

it is difficult for us to infer the type of named entity Rob. It may describe a person or an animal. With the help of its accompanying image， we can easily determine that its type is MISC （other）.

There are four types of entities: Person (PER),Organization (ORG), Location (LOC) and others (MISC).

Current research and existing problems:

They mainly focus on using a cross-modal attention mechanism to combine text representation with image representation.

- the current methods are based on a strong assumption that each text and its accompanying image are matched, and the image can be used to help identify named entities in the text.
- the current methods fail to construct a consistent representation to bridge the semantic gap between two modalities, which prevents the model from establishing a good connection between the text and image.

To address these issues:

propose a general matching and alignment framework (MAF)

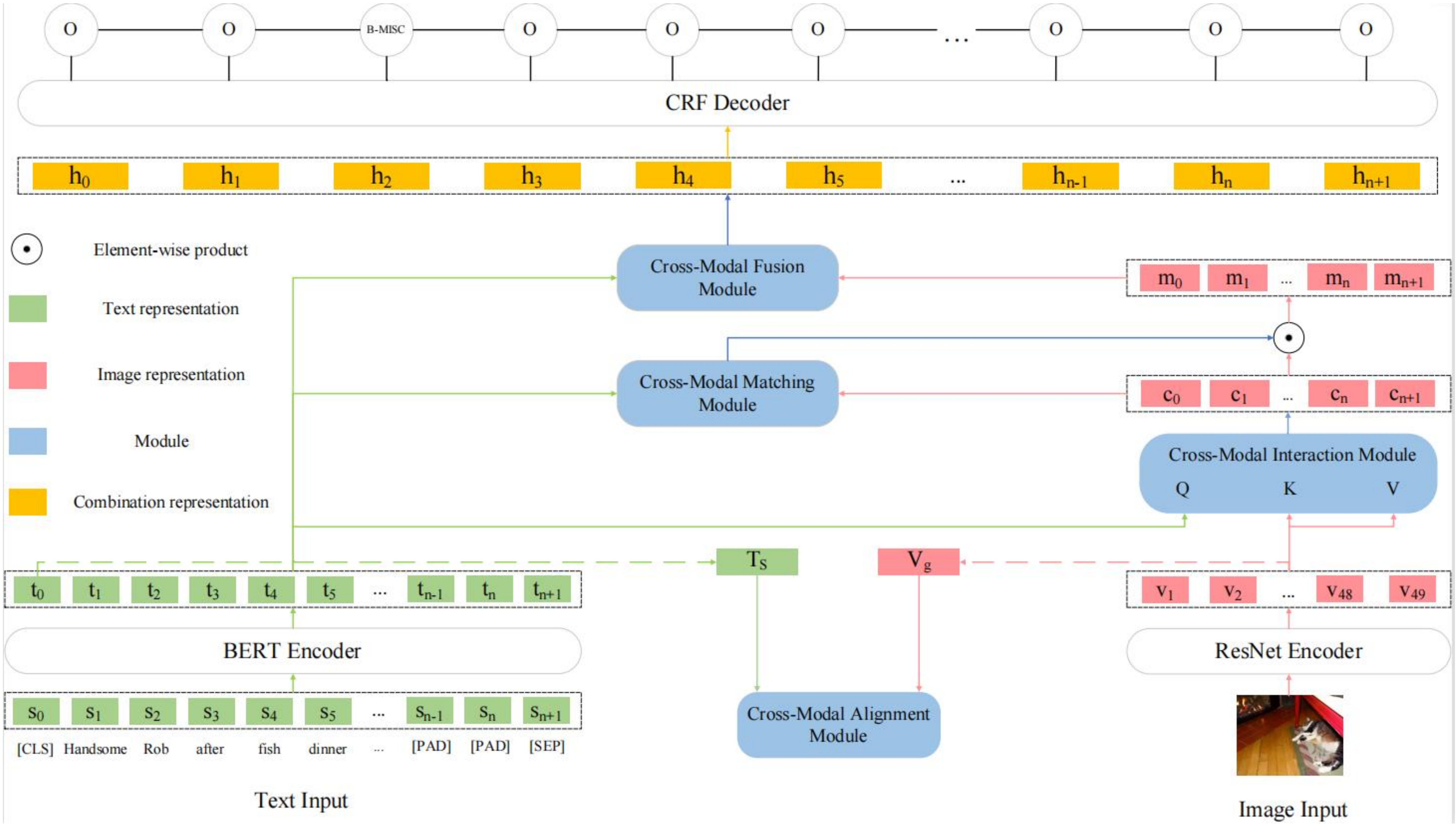
a cross-modal matching (CM) module：

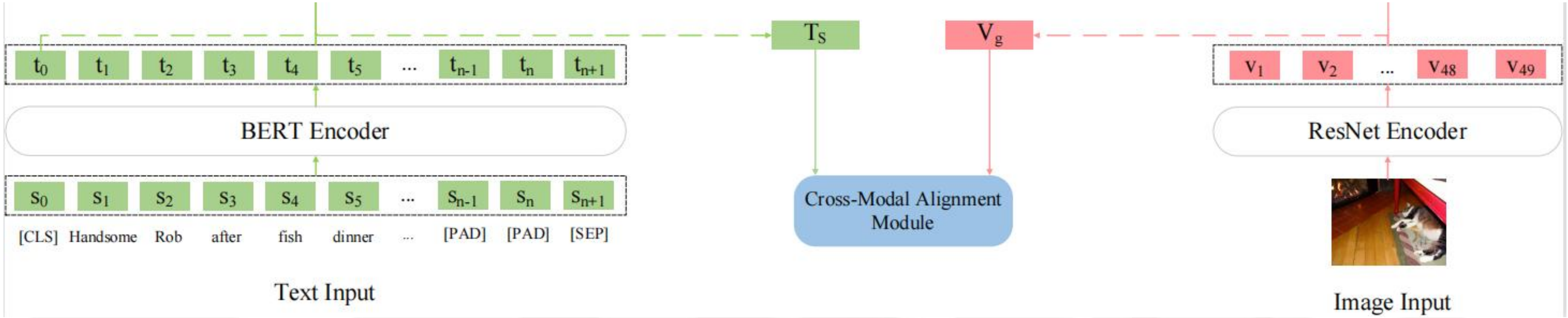
—— reduce the impact of mismatched text-image pairs.

a cross-modal alignment (CA) module：

—— help the model to align the text and image representations.

概观 OVERVIEW


O
B-MISC
CRF Decoder
h0 h1 h2 h3 h4 h5 ... hn-1 hn hn+1
Element-wise product
Text representation
Image representation
Module
Combination representation
Cross-Modal Fusion Module
m0 m1 ... mn mn+1
Cross-Modal Matching Module
c0 c1 ... cn cn+1
Cross-Modal Interaction Module
Q K V
Ts
Vg
t0 t1 t2 t3 t4 t5 ... tn-1 tn tn+1
v1 v2 ... v48 v49
BERT Encoder
ResNet Encoder
S0 S1 S2 S3 S4 S5 ... Sn-1 Sn Sn+1
[CLS] Handsome Rob after fish dinner ... [PAD] [PAD] [SEP]
Cross-Modal Alignment Module
Text Input
Image Input



Input Representations

- Text Encoder

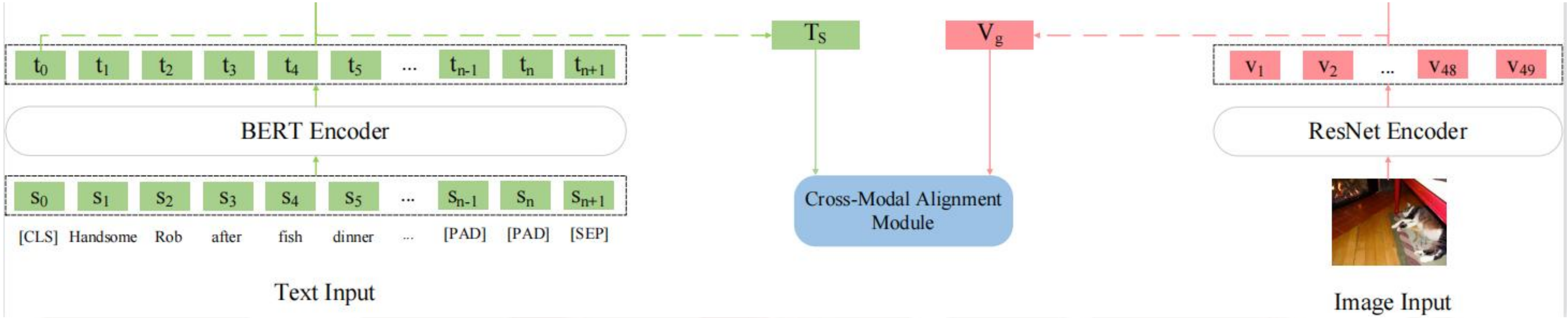
$$S' = (s_0, s_1, \ldots, s_n, s_{n+1})$$

$$T = (t_0, t_1, \ldots, t_n, t_{n+1})$$

$$T_s = Tanh(t_0) = \frac{e^{2t_0} - 1}{e^{2t_0} + 1}$$

- Image Encoder

$$V = (v_1, v_2, \ldots, v_{48}, v_{49})$$



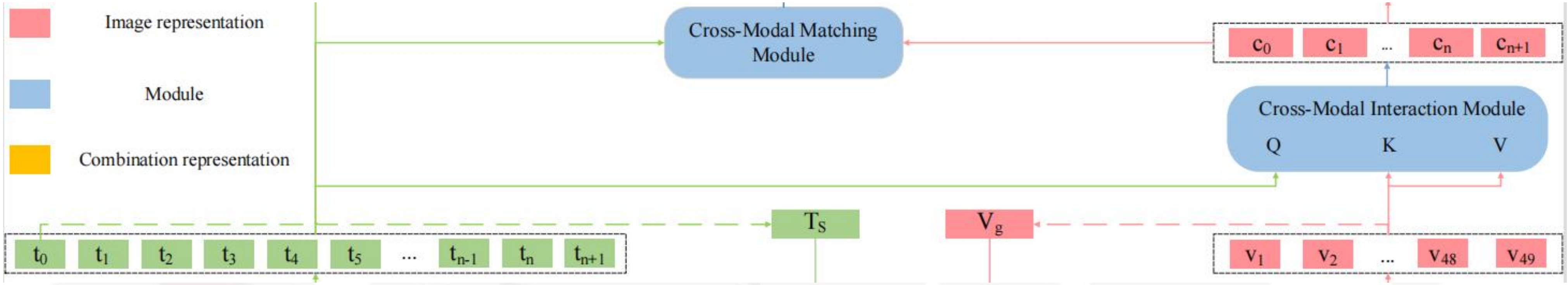
Cross-Modal Alignment Module

- the effect of contrastive learning is mainly affected by the number of negative examples.
- this MLP projection can help the encoders (BERT and ResNet) to learn a better representation.
- By minimizing two contrast loss functions, we can maximize the similarity of positive cases and minimize the similarity of negative cases.

$$\mathcal{L}_i^{(Vc\to Tc)} = -\log \frac{e^{(sim(V_c^i, T_c^i)/\tau)}}{\sum_{j=1}^{N} e^{(sim(V_c^i, T_c^j)/\tau)}}$$

$$\mathcal{L}_i^{(Tc\to Vc)} = -\log \frac{e^{(sim(T_c^i, V_c^i)/\tau)}}{\sum_{j=1}^{N} e^{(sim(T_c^i, V_c^j)/\tau)}}$$

$$\mathcal{L}_{ca} = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda_c \mathcal{L}_i^{(Vc\to Tc)} + (1-\lambda_c)\mathcal{L}_i^{(Tc\to Vc)}\right)$$



Cross-Modal Interaction Module

- **Queries**

$$T = (t_0, t_1, \ldots, t_n, t_{n+1})$$

- **Key-value pairs**

$$V = (\mathrm{v}_1, v_2, \ldots, v_{48}, v_{49})$$

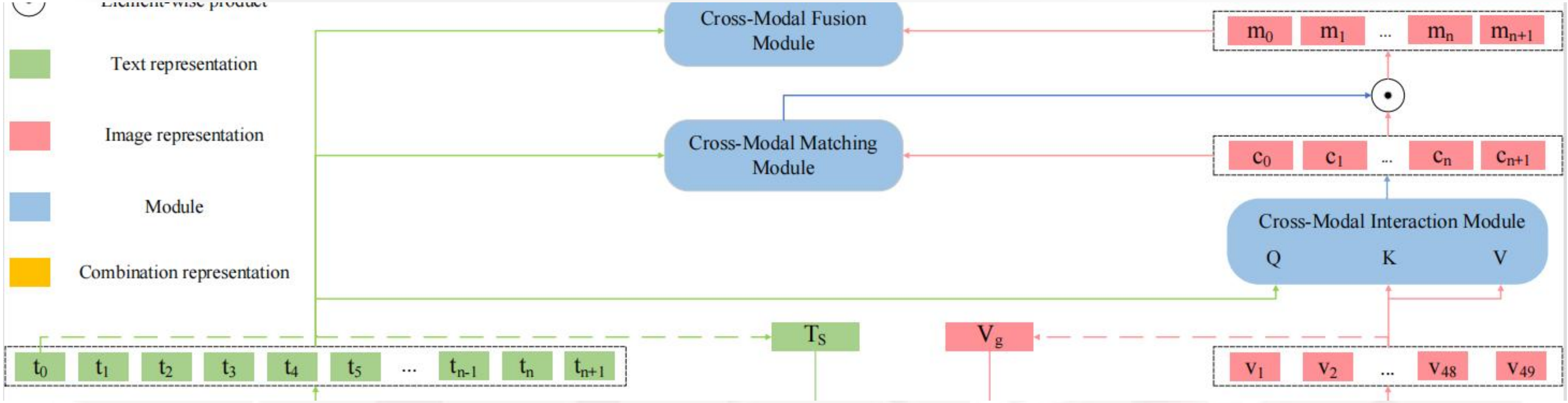
$$a_i = \mathrm{softmax}\left(\frac{[W_{qi}T]^T[W_{ki}V]}{\sqrt{d/m}}\right)$$

$$CV_i = a_i\ [W_{vi}T]^T$$

$$CV = W'\ [CV_1; CV_2; \ldots; CV_m]^T$$

$$V' = LN(T + CV)$$

$$C = LN(V' + FFN(V'))$$



Cross-Modal Matching Module

- Randomly select $2k$ $(0 < k < N/2)$ input pairs from the batch and swap the image representations of the first half in the input pairs with the second half as the negative examples. Moreover, the remaining $N - 2k$ input pairs in the batch are positive examples.
- Use the generated training example to train the CM module.

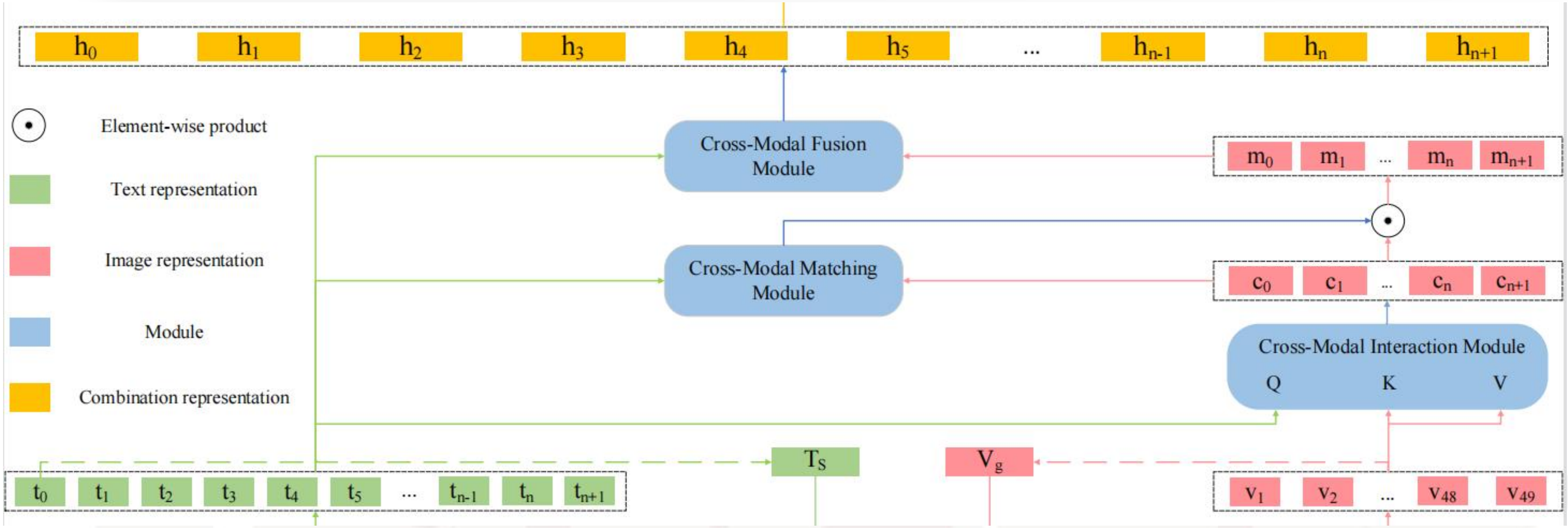
$$D_m = (D_{m1}, D_{m2}, \ldots, D_{mN})$$

$$F_i = Flatten([T(D_{mi}); C(D_{mi})])$$

$$\hat{y}_{mi} = \sigma(W_f^T F_i)$$

$$\mathcal{L}_{cm} = -\frac{1}{N}\sum_{j=1}^{N}\left(y_{mj} \cdot \log(\hat{y}_{mj}) + (1 - y_{mj}) \cdot \log(1 - \hat{y}_{mj})\right)$$

$$M = \hat{y}_m \odot C$$



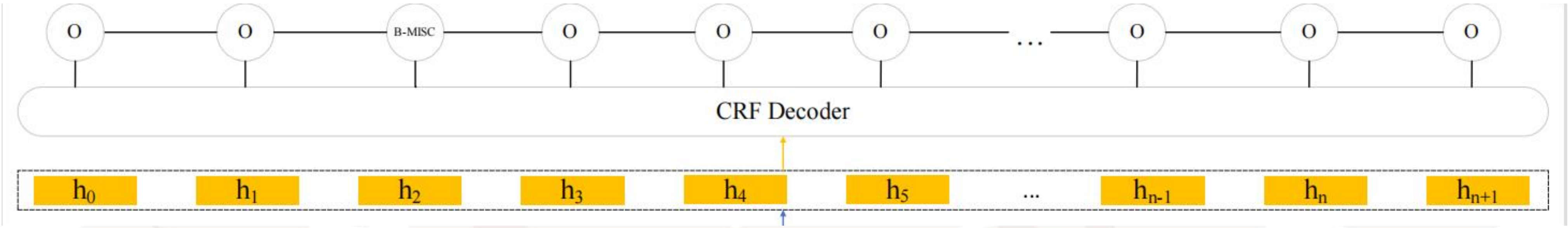
Cross-Modal Fusion Module

- use a gate mechanism to dynamically control the combination of text and image representations at the token level.

$$g = \sigma(W_{gt}^T T + W_{gm}^T M)$$

$$R = g \odot M$$

$$\mathrm{H} = [T; R]$$



CRF Decoder

$$P(y|S,I)=\frac{e^{(\sum_{i=1}^{n}E_{h_i,y_i}+\sum_{i=0}^{n}T_{y_i,y_{i+1}})}}{Z(H)}$$

$$\mathcal{L}_{mner}=-\frac{1}{|D_{mner}|}\sum_{j=1}^{N}(\log P(y^j|S^j,I^j))$$

$$Z(H)=\sum_{y}e^{(\sum_{i=1}^{n}E_{h_i,y_i}+\sum_{i=0}^{n}T_{y_i,y_{i+1}})}$$

$$\mathcal{L}=\alpha\mathcal{L}_{ca}+\beta\mathcal{L}_{cm}+(1-\alpha-\beta)\mathcal{L}_{mner}$$

Type	TWITTER-2015 Train	TWITTER-2015 Dev	TWITTER-2015 Test	TWITTER-2017 Train	TWITTER-2017 Dev	TWITTER-2017 Test
PER	2,217	552	1,816	2,943	626	621
LOC	2,091	522	1,697	731	173	178
ORG	928	247	839	1,674	375	395
MISC	940	225	726	701	150	157
Total	6,176	1,546	5,078	6,049	1,324	1,351
# Tweets	4,000	1,000	3,257	3,373	723	723

实验 Experiments

Methods	TWITTER-2015							TWITTER-2017						
	Single Type (F1)				Overall			Single Type (F1)				Overall		
	PER.	LOC.	ORG.	MISC.	P	R	F1	PER.	LOC.	ORG.	MISC.	P	R	F1
BiLSTM-CRF	76.77	72.56	41.33	26.80	68.14	61.09	64.42	85.12	72.68	72.50	52.56	79.42	73.43	76.31
CNN-BiLSTM-CRF	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80.00	78.76	79.37
HBiLSTM-CRF	82.34	76.83	51.59	32.52	70.32	68.05	69.17	87.91	78.57	76.67	59.32	82.69	78.16	80.37
BERT	84.72	79.91	58.26	38.81	68.30	74.61	71.32	90.88	84.00	79.25	61.63	82.19	83.72	82.95
BERT-CRF♣	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
T-NER♠	83.64	76.18	50.26	34.56	69.54	68.65	69.09	-	-	-	-	-	-	-
GVATT-HBiLSTM-CRF	82.66	77.21	55.06	35.25	73.96	67.90	70.80	89.34	78.53	79.12	62.21	83.41	80.38	81.87
AdaCAN-CNN-BiLSTM-CRF	81.98	78.95	53.07	34.02	72.75	68.74	70.69	89.63	77.46	79.24	62.77	84.16	80.24	82.15
GVATT-BERT-CRF	84.43	80.87	59.02	38.14	69.15	74.46	71.70	90.94	83.52	81.91	62.75	83.64	84.38	84.01
AdaCAN-BERT-CRF	85.28	80.64	59.39	38.88	69.87	74.59	72.15	90.20	82.97	82.67	64.83	85.13	83.20	84.10
MT-BERT-CRF	**85.30**	81.21	61.10	37.97	70.48	74.80	72.58	91.47	82.05	81.84	65.80	84.60	84.16	84.42
UMT-BERT-CRF♣	85.24	**81.58**	63.03	39.45	71.67	**75.23**	73.41	**91.56**	84.73	82.24	**70.10**	85.28	85.34	85.31
ATTR-MMKG-MNER♠	84.28	79.43	58.97	41.47	**74.78**	71.82	73.27	-	-	-	-	-	-	-
MAF (Ours)	84.67	81.18	**63.35**	**41.82**	71.86	75.10	**73.42**	91.51	**85.80**	**85.10**	68.79	**86.13**	**86.38**	**86.25**

Methods	TWITTER-2015		TWITTER-2017		**Size (M)**
	Training	**Testing**	**Training**	**Testing**	
UMT-BERT-CRF	102.035	30.002	85.971	6.281	208.29
MAF	**86.822**	**25.619**	**73.754**	**5.450**	**196.28**

Methods	TWITTER-2015			TWITTER-2017		
	P	R	F1	P	R	F1
MAF	**71.41**	75.32	**73.32**	**86.13**	**86.38**	**86.25**
w/o CA	70.89	**75.44**	73.09	83.75	84.68	84.21
w/o CM	70.96	74.73	72.80	85.40	84.46	84.93
w/o CA + CM	70.32	74.71	72.45	82.90	84.30	83.60

实验 Experiments

Methods	Importance of the *CA* Module		Importance of the *CM* Module	
				
	[**HURRY O**] GET ONE BEFORE THEYRE SENT TO AFRICA	The beautiful camel is called [**Camille MISC**]	[**Aquamarine MISC**] (2006)	#[**Malevich PER**] opens at Tate Modern on 16 July
UMT-BERT-CRF	[**HURRY PER**] ×	[**Camille MISC**] ✓	[**Aquamarine ORG**] ×	[**Malevich PER**] ✓
MAF	[**HURRY O**] ✓	[**Camille MISC**] ✓	[**Aquamarine MISC**] ✓	[**Malevich PER**] ✓
MAF w/o CA	[**HURRY PER**] ×	[**Camille PER**] ×	[**Aquamarine MISC**] ✓	[**Malevich PER**] ✓
MAF w/o CM	[**HURRY O**] ✓	[**Camille MISC**] ✓	[**Aquamarine ORG**] ×	[**Malevich LOC**] ×

致谢感恩

孙凯文 | 导师：杨武 黄丹

THANKS TO