

# Learning from Different text-image Pairs: A Relation-enhanced Graph Convolutional Network for Multimodal NER

Fei Zhao, Chunhui Li, Zhen Wu, Shangyu Xing, Xinyu Dai



# Outline

---



- Introduction
- Motivation
- Methodology
- Experiments
- Conclusions

# Outline

---



- Introduction
- Motivation
- Methodology
- Experiments
- Conclusions

# Named Entity Recognition

---



- Named Entity Recognition (NER) is a subtask of information extraction, which aims to identify text spans to specific entity types such as Person (PER), Location (LOC) and Organization (ORG)
  - entity linking [1]
  - relation extraction [2]

# Multimodal Named Entity Recognition



- the texts contain polysemy entities



Figure 1: An example of multimodal tweets. In this tweet, “Alibaba” is the name of a Person instead of an Organization.

- aligning the words in the text with the visual objects in the image
  - encoding the whole image into a global feature vector [3];
  - segmenting the whole image averagely into multiple visual regions [4,5,6,7,8,9];
  - only retaining the visual object regions in the image [10,11,12];

# Outline

---



- Introduction
- **Motivation**
- Methodology
- Experiments
- Conclusions

# external matching relations



- Inter-modal relation

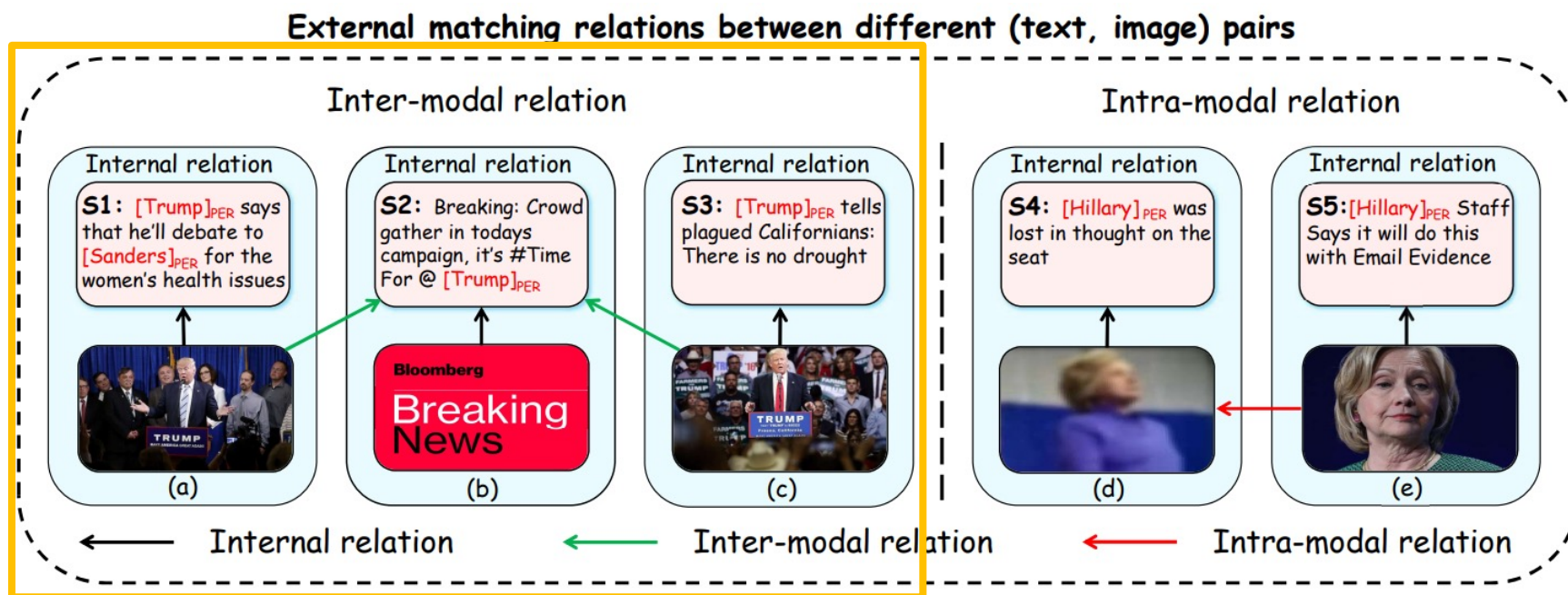


Figure 2: Each blue box contains a pair of image and text in the dataset. The black arrow represents the internal matching relation in a image-text pair. The green arrow represents the inter-modal relation between the text and image in different image-text pairs.



# external matching relations



- Intra-modal relation

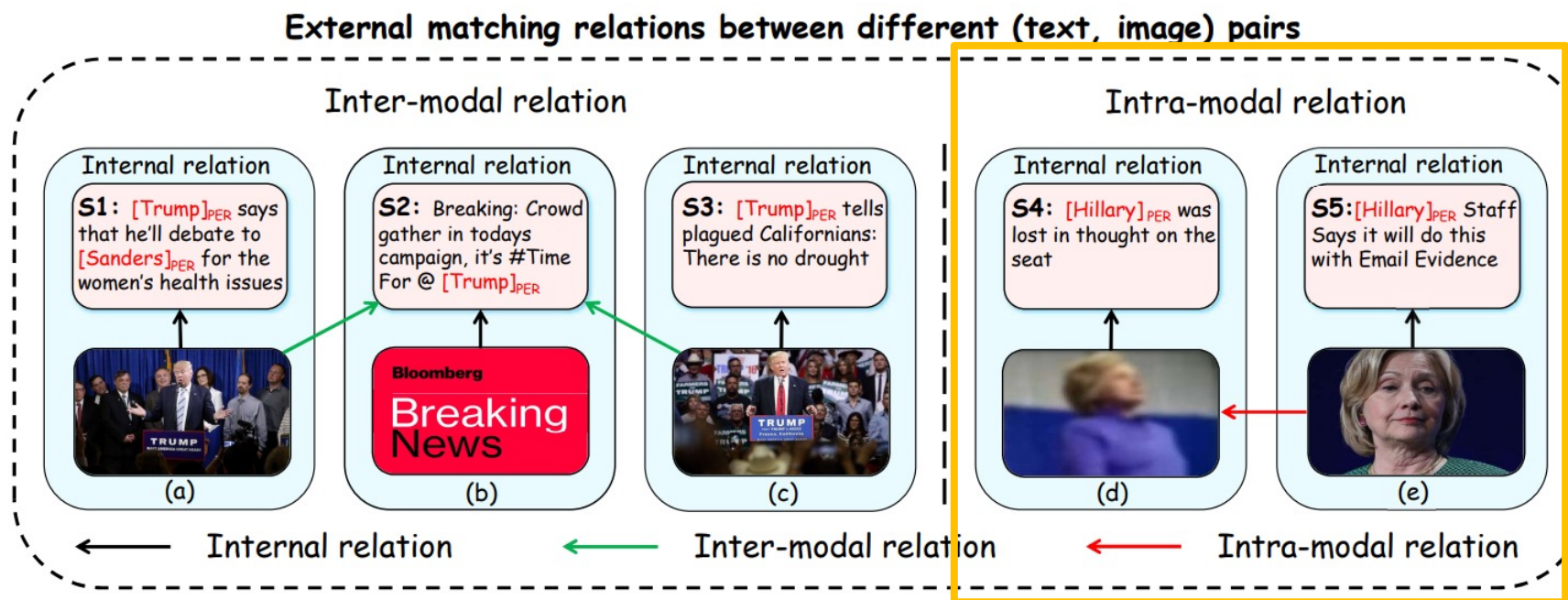


Figure 2: Each blue box contains a pair of image and text in the dataset. The black arrow represents the internal matching relation in a image-text pair. The red arrow represents the intra-modal relation between images in different image-text pairs.

# Outline

---



- Introduction
- Motivation
- **Methodology**
- Experiments
- Conclusions

# The Proposed Model

- model two kinds of external matching relations

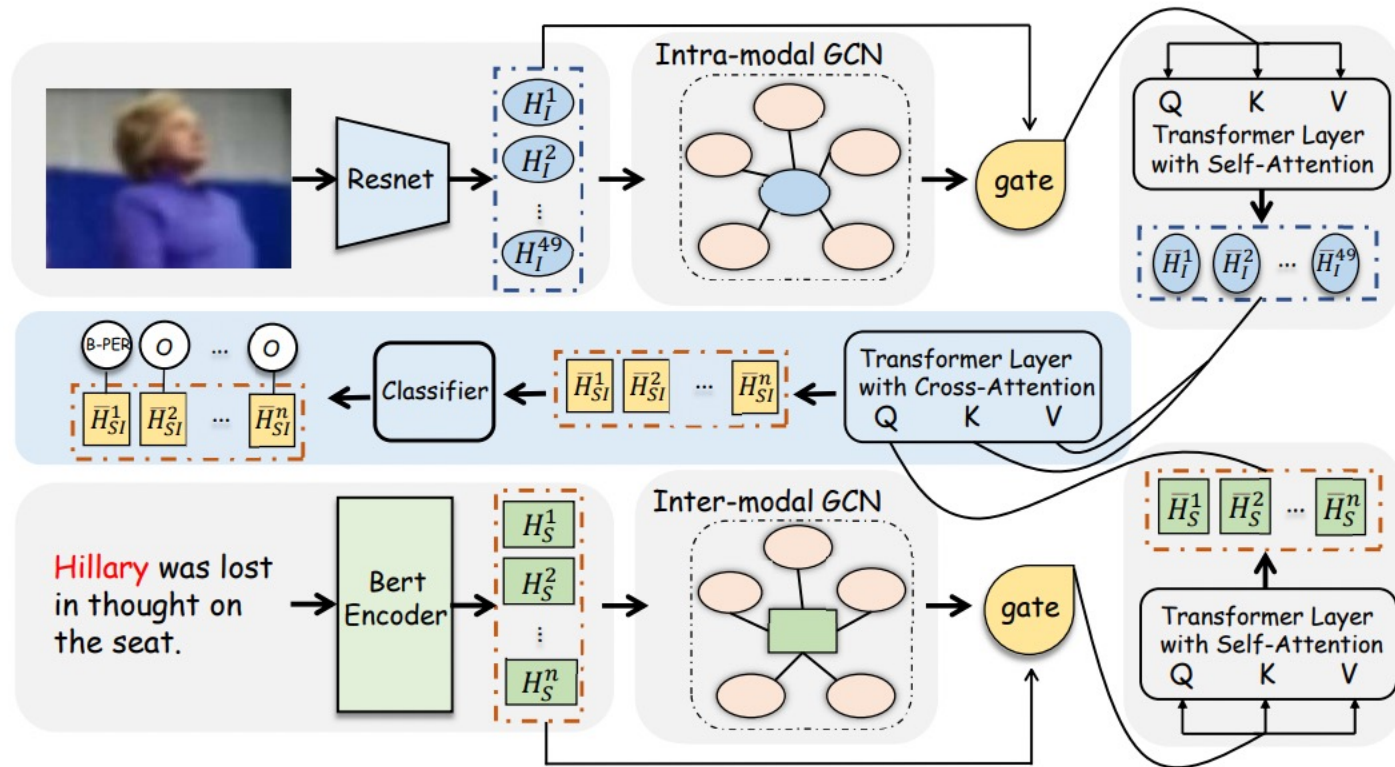
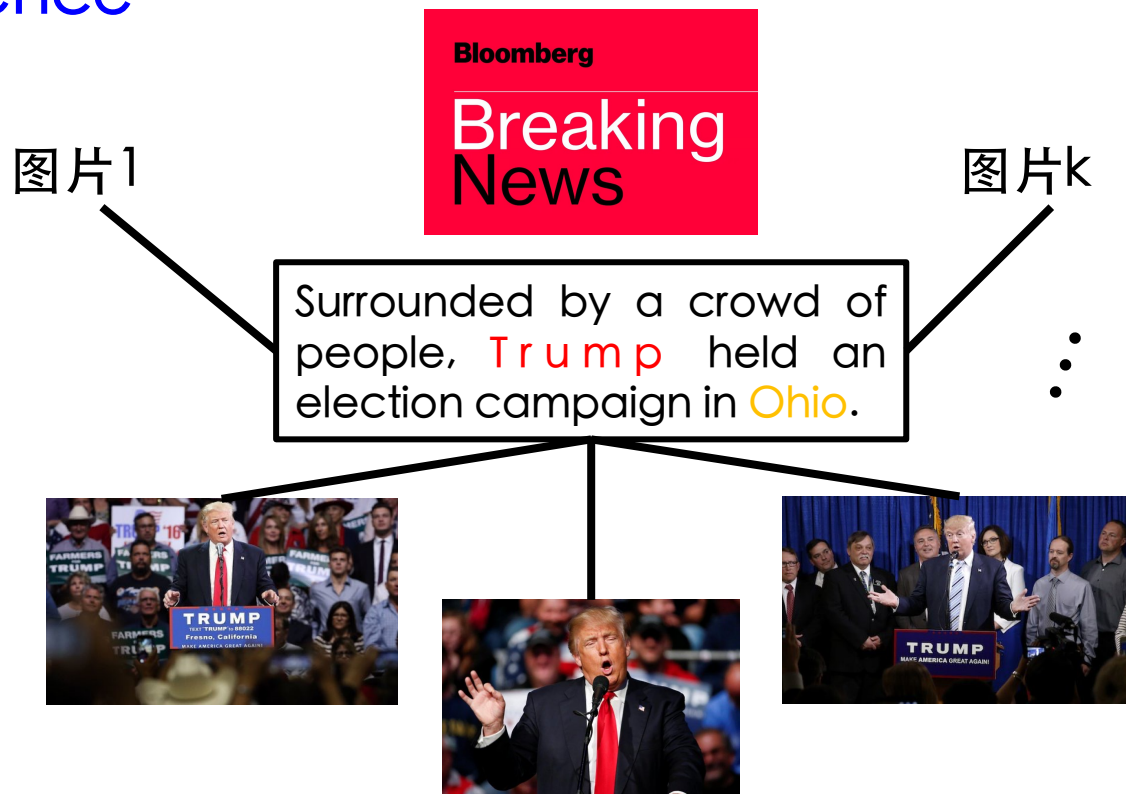


Figure 3: The overview of our proposed Relation-enhanced Graph Convolutional Network (R-GCN).

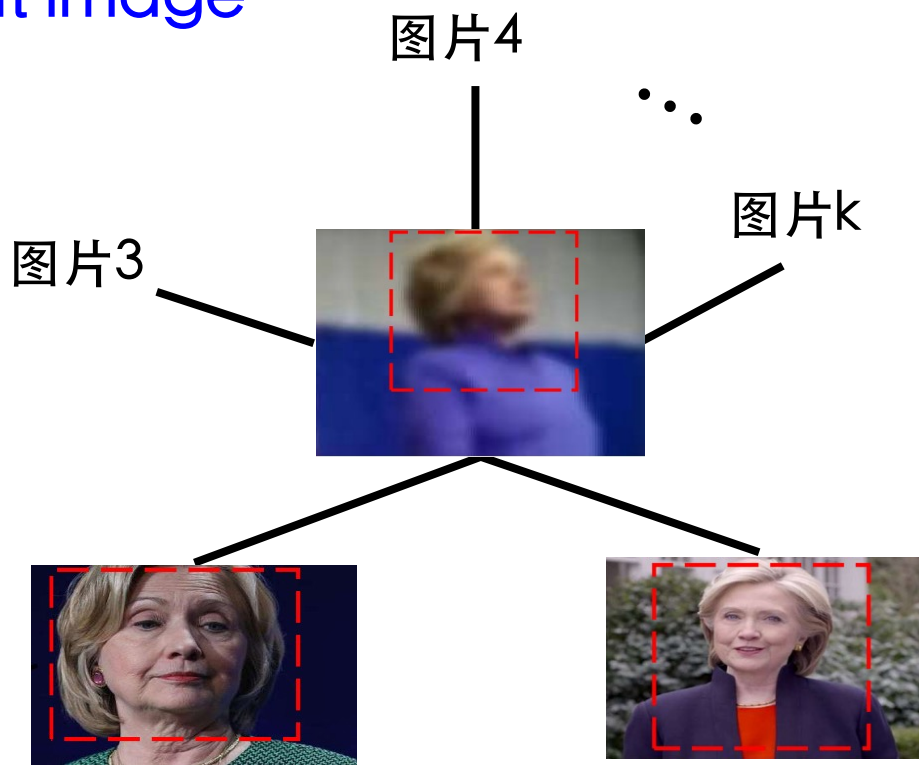
# Inter-Modal Relation Module

- Nodes: text node and image node
- Edges: measure whether other images in the dataset contain similar scenes mentioned in the sentence



# Intra-Modal Relation Module

- Nodes: image node
- Edges: measure whether other images in the dataset contain the same types of visual object with input image



# Outline

---



- Introduction
- Motivation
- Methodology
- **Experiments**
- Conclusions



# Main Results



Methods	TWITTER-2015							TWITTER-2017						
	Single Type ( $F_1$ )				Overall			Single Type ( $F_1$ )				Overall		
	PER	LOC	ORG	MISC	Pre	Rec	F1	PER	LOC	ORG	MISC	Pre	Rec	F1
<i>Text</i>														
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
<i>Text+Image</i>														
ACOA	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
VG	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
OCSGA	84.68	79.95	56.64	39.47	74.71	71.21	72.92	–	–	–	–	–	–	–
Object-AGBAN	84.75	79.41	58.31	40.72	74.13	72.39	73.25	–	–	–	–	–	–	–
IAIK	84.28	79.42	58.97	41.47	<b>74.78</b>	71.82	73.27	–	–	–	–	–	–	–
RpBERT*	85.18	81.19	58.68	37.88	71.15	74.30	72.69	89.05	84.03	82.60	63.67	82.85	84.38	83.61
UMT	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
UMT*	84.74	81.69	60.59	39.22	72.66	74.14	73.39	90.41	83.98	81.20	65.56	84.02	84.09	84.05
UMGF	84.26	<b>83.17</b>	<b>62.45</b>	<b>42.42</b>	74.49	75.21	74.85	91.92	85.22	83.13	69.83	86.54	84.50	85.51
UMGF*	84.50	81.54	60.72	40.57	72.47	74.60	73.52	91.14	84.24	83.23	67.30	85.30	84.99	85.14
<b>R-GCN</b>	<b>86.36</b>	82.08	60.78	41.56	73.95	76.18	<b>75.00</b> <sup>†</sup>	<b>92.86</b>	<b>86.10</b>	<b>84.05</b>	<b>72.38</b>	<b>86.72</b>	87.53	<b>87.11</b> <sup>†</sup>
	$\pm 0.31$	$\pm 0.21$	$\pm 0.64$	$\pm 0.86$	$\pm 0.32$	$\pm 0.53$	$\pm 0.18$	$\pm 0.46$	$\pm 0.94$	$\pm 0.74$	$\pm 1.79$	$\pm 0.45$	$\pm 0.34$	$\pm 0.36$
<b>R-GCN (w/o Gate)</b>	86.10	81.90	60.17	41.59	72.50	<b>76.89</b>	74.60	92.74	85.89	83.01	72.35	85.90	<b>87.57</b>	86.70
	$\pm 0.37$	$\pm 1.74$	$\pm 0.48$	$\pm 0.99$	$\pm 0.79$	$\pm 0.79$	$\pm 0.36$	$\pm 0.86$	$\pm 1.09$	$\pm 1.07$	$\pm 1.72$	$\pm 1.19$	$\pm 0.49$	$\pm 0.52$

Table 1: Performance comparison on the TWITTER-15 and TWITTER-17 datasets (%).

# Ablation Study



Methods	TWITTER-2015							TWITTER-2017						
	Single Type ( $F_1$ )				Overall			Single Type ( $F_1$ )				Overall		
	PER	LOC	ORG	MISC	Pre	Rec	F1	PER	LOC	ORG	MISC	Pre	Rec	F1
R-GCN	86.36	82.08	60.78	41.56	73.95	76.18	75.00	92.86	86.10	84.05	72.38	86.72	87.53	87.11
w/o InterRG	85.52	81.16	59.30	40.74	73.42	74.79	74.05	92.63	85.32	81.55	72.29	85.34	87.07	86.17
w/o IntraRG	85.41	81.75	60.66	40.01	73.15	75.55	74.29	92.90	82.58	82.73	71.82	85.44	87.05	86.22
w/o InterRG, IntraRG	85.51	81.45	58.72	37.61	73.18	74.09	73.59	93.58	81.59	80.12	71.37	84.12	86.21	85.13

Table 2: Ablation study over two main components of proposed model (%).



# Case Study

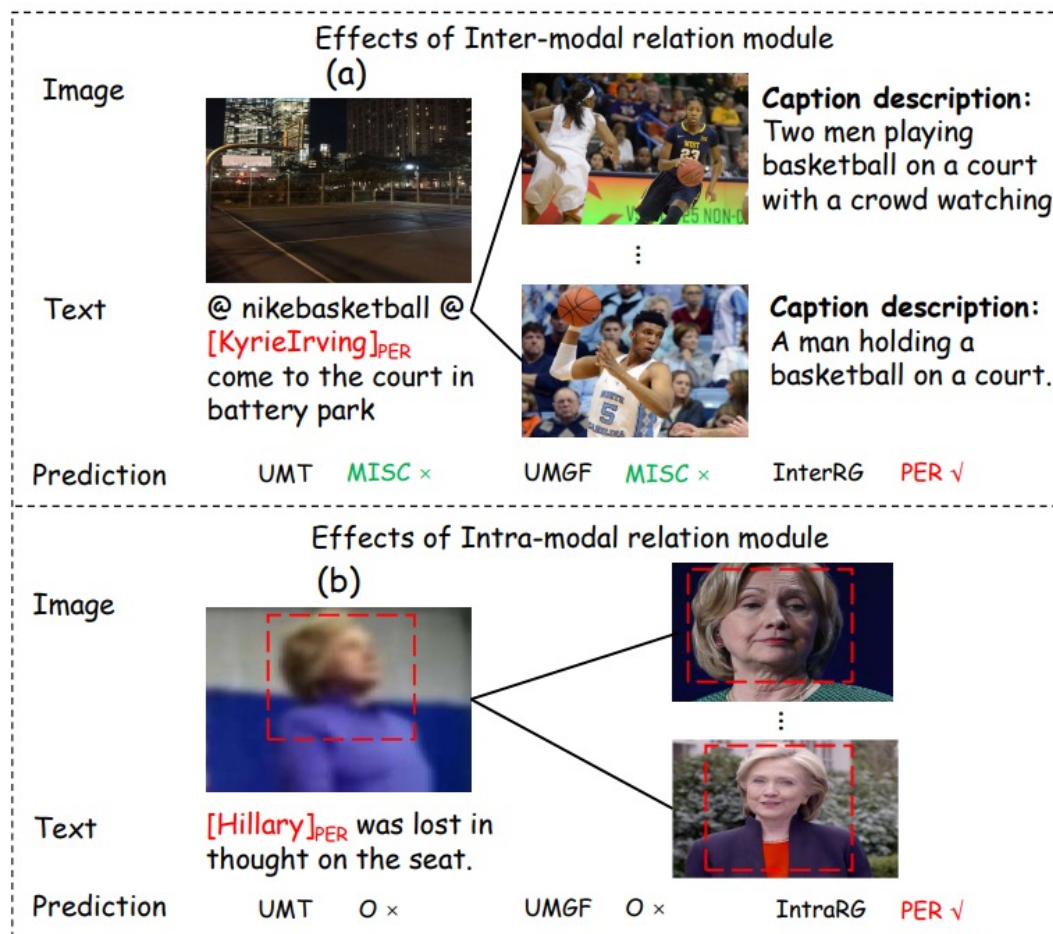


Figure 4: Predictions of UMT, UMGF, InterRG and IntraRG on two test samples. **×** and **✓** denote incorrect and correct predictions.

# Error Analysis



- bias brought by annotation
- lack of background knowledge
- information deficiency




	(a)	(b)	(c)
Image			
Text	RT @1Kindesign : [Pebble Beach Residence] <sub>ORG</sub> with luxury spa ambiance	Forever my favorite [Jonas brother] <sub>ORG</sub>	[Welkom] <sub>LOC</sub> in 1992
Error Type	Bias brought by annotation (40%)	Lack of background knowledge (22%)	Information Deficiency (10%)
Ground Truth	ORG ✓	ORG ✓	LOC ✓
R-GCN	LOC ×	PER ×	ORG ×

Figure 5: Three typical errors of R-GCN.

# Outline

---



- Introduction
- Motivation
- Methodology
- Experiments
- **Conclusions**

- we propose a novel Relation-enhanced Graph Convolutional Network for the Multimodal Named Entity Recognition task
- The main idea of our approach is to leverage two kinds of external matching relations in different (image, text) pairs to improve the ability of identifying named entities in the text
- Results from numerous experiments indicate that our model achieves better performance than other state-of-the-art methods.

Thank You

- [1] Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep Joint Entity Disambiguation with Local Neural Attention. In EMNLP 2017.
- [2] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. Kernel Methods for Relation Extraction. In EMNLP 2002.
- [3] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual Attention Model for Name Tagging in Multimodal Social Media. In Proc. of ACL.
- [4] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal Named Entity Recognition for Short Social Media Posts. In Proc. of NAACL.
- [5] Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen. 2021. Multimodal Named Entity Recognition with Image Attributes and Image Knowledge. In Proc. of DAS-FAA.

# References

---



- [6] Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020. RIVA: A Pre-trained Tweet Multimodal Model Based on Text-image Relation for Multimodal NER. In Proc. of COLING.
- [7] Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rp-BERT: A Text-image Relation Propagation-based BERT Model for Multimodal NER. In Proc. of AAAI.
- [8] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In Proc. of ACL.
- [9] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive Co-attention Network for Named Entity Recognition in Tweets. In Proc. of AAAI.

# References

---



- [10] Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. 2020. Multimodal Aspect Extraction with Region-Aware Alignment Network. In Proc. of NLPCC
- [11] Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and QingLi. 2020. Multimodal Representation with Embedded Visual Guiding Objects for Named Entity Recognition in Social Media Posts. In Proc. of ACM MM.
- [12] Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance. In Proc. of AAAI.