Information Screening whilst Exploiting! Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling

Shengqiong Wu[1], Hao Fei[1], Yixin Cao[2], Lidong Bing[3], Tat-Seng Chua[1]

[1] Sea-NExT Joint Lab, School of Computing, National University of Singapore
[2] Singapore Management University, [3] DAMO Academy, Alibaba Group

Paper

A. Motivation

▶ **Internal-information over-utilization.** Prior research shows that only parts of the texts are useful to the relation inference, and not all and always the visual sources play positive roles for MRE. A fine-grained feature screening over both the internal image and text features is needed.

▶ **External-information under-exploitation.** Although compensating the texts with visual sources, there can be still information deficiency in MRE, in particular when the visual features serve less (or even negative) utility. More external semantic supplementary information should be exploited for MRE.

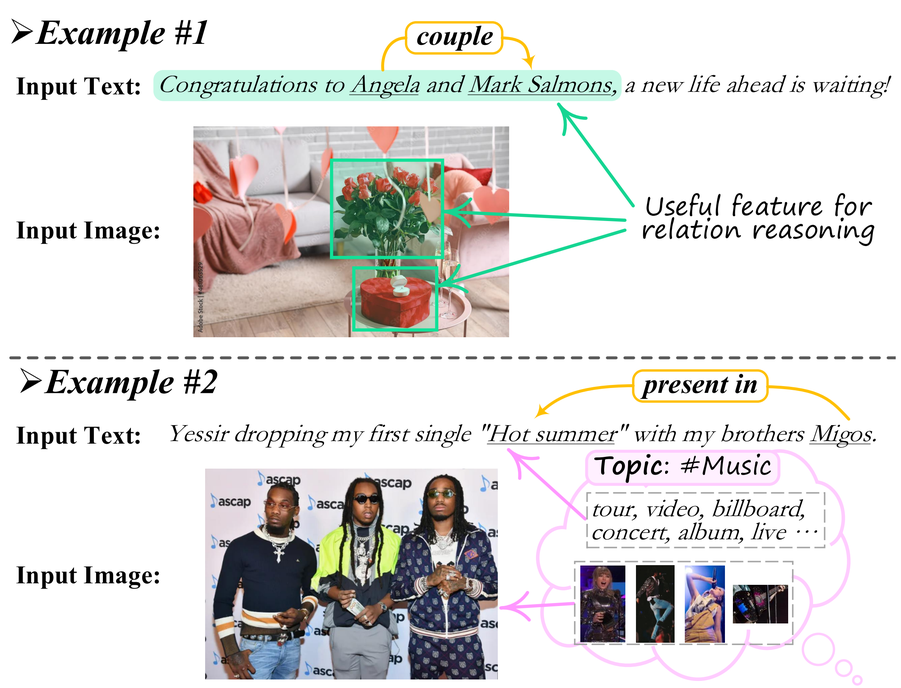

Figure 1 – Examples of multimodal relation extraction (MRE). The relational pairs are marked in texts.

B. Method

As shown in Figure 2, our overall framework consists of five tiers:

★ **Scene Graph Generation.** The model takes as input an image I and text T, as well as the subject v_s and object entity v_o. We represent I and T with the corresponding visual scene graph (VSG) and textual scene graph (TSG).

★ **Cross-modal Graph Construction.** The VSG and TSG are assembled as a cross-modal graph (CMG), which is further modeled via a graph encoder.

★ **GIB-guided Feature Refinement.** We perform GIB-guided feature refinement (GENE) over the CMG for internal-information screening, i.e., node filtering and edge adjusting, which results in a structurally compact backbone graph.

★ **Multimodal Topic Integration.** The multimodal topic features induced from the latent multimodal topic model (LAMO) are integrated into the previously obtained compressed feature representation for external-information exploitation via an attention operation.

★ **Inference.** The decoder predicts the relation label Y based on the enriched features.

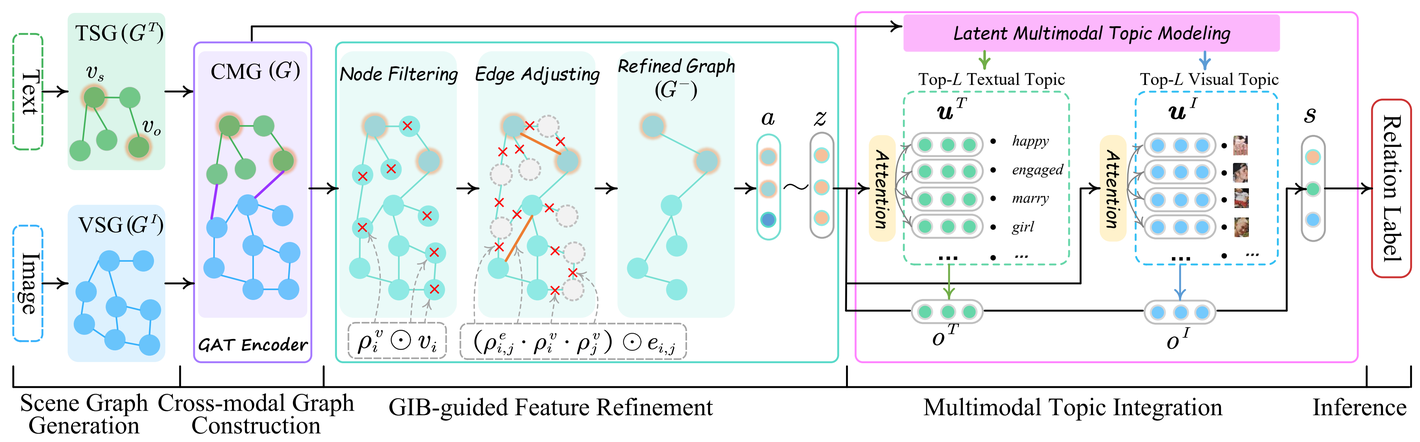

Figure 2 – Overview of our proposed framework.

C. Main Results

▶ Multimodal methods, by leveraging the additional visual features, exhibit higher performances consistently.
▶ Our model boosts the SoTA with a very significant margin.
▶ Information screening and exploiting both contribute to task performance improvements.
▶ The scene graph is beneficial for the structural modeling of the multimodal inputs.

	Acc.	Pre.	Rec.	F1
• ***Text-based Methods***				
BERT†	-	63.85	55.79	59.55
PCNN†	72.67	62.85	49.69	55.49
MTB†	72.73	64.46	57.81	60.86
DP-GCN♭	74.60	64.04	58.44	61.11
• ***Multimodal Methods***				
BERT(Text+Image)♭	74.59	63.07	59.53	61.25
BERT+SG†	74.09	62.95	62.65	62.80
MEGA†	76.15	64.51	68.44	66.41
VisualBERT†$_{base}$	-	57.15	59.48	58.30
ViLBERT†$_{base}$	-	64.50	61.86	63.16
RDS†	-	66.83	65.47	66.14
HVPNeT†	-	83.64	80.78	81.85
MKGformer†	92.31	82.67	81.25	81.95
Ours	**94.06**	**84.69**	**83.38**	**84.03**
w/o GENE	92.42	82.41	81.83	82.12
w/o $I(\boldsymbol{z},G)$	93.64	83.61	82.34	82.97
w/o LAMO	92.86	82.97	81.22	82.09
w/o $\boldsymbol{o}^T$	93.05	83.95	82.53	83.23
w/o $\boldsymbol{o}^I$	93.63	84.03	83.18	83.60
w/o VSG&TSG	93.12	83.51	82.67	83.09
w/o CMG	93.97	84.38	83.20	83.78

Table 1 – Main Results. 'w/o $I(\boldsymbol{z},G)$' means GENE adjustment without GIB guidance. 'w/o CMG' means VSG and TSG are not connected with hyper-edge $E^{\times}$. 'w/o VSG&TSG' means our method uses the embedding of visual and text inputs without structural SG modeling. Baselines with the superscript '†' are copied from the raw papers; with '♭' are from our re-implementation.

D. In-depth Analysis

▶ **RQ1**: *Does* GENE *helps by really denoising the input features?*. **A**: Yes, cf. Figure 3.
▶ **RQ2**: *Are* LAMO *induced task-relevant topic features beneficial to the end task?* **A**: Yes, cf. Figure 5 & Figure 6.
▶ **RQ3**: *How do* GENE *and* LAMO *collaborate to solve the end task?* **A**: cf. Figure 7.
▶ **RQ4**: *Under what circumstances do the internal-information screening and external-information exploiting help?* **A**: cf. Figure 4.

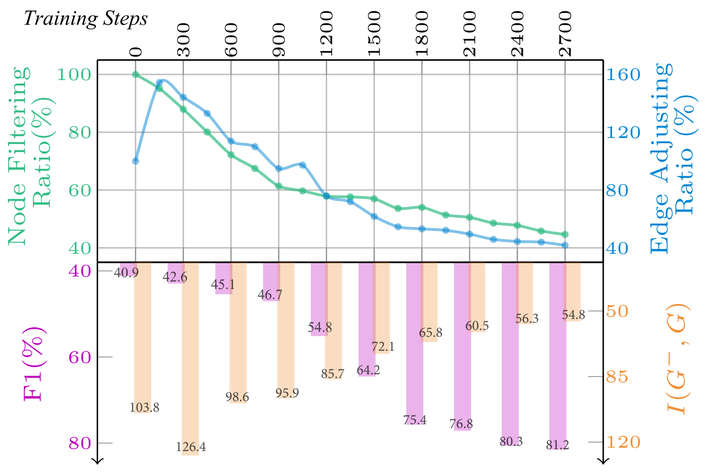

Figure 3 – The trends of changing ratio of nodes and edges, along with the task performance and the mutual information between G and G^-. The model is without LAMO.

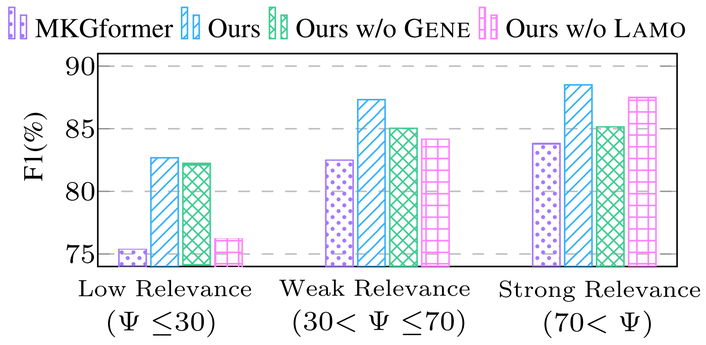

Figure 4 – Results under varying text-image relevance.

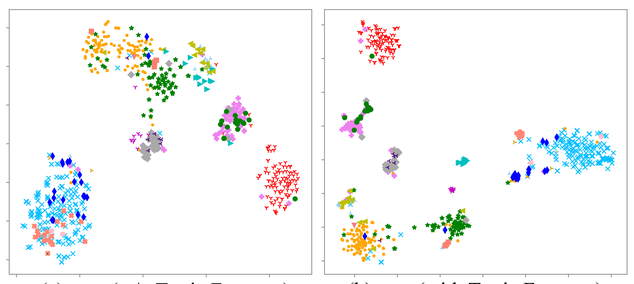

Figure 5 – T-SNE visualization of the contextualized features with or without topic features. Different colors indicate different ground-truth relation labels.

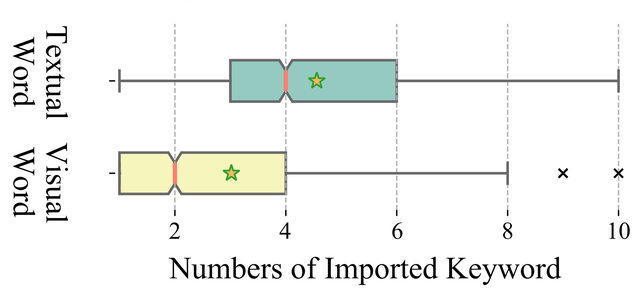

Figure 6 – Distribution of numbers of textual and visual topic keywords imported for MRE.

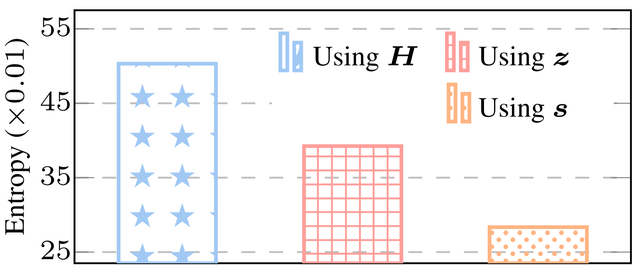

Figure 7 – The entropy under various model settings.