Learning Semantic Relationship Among Instances for Image-Text Matching

Zheren Fu · Zhendong Mao · Yan Song · Yongdong Zhang

University of Science and Technology of China, Hefei, China

Image-text Matching

- ✓ Image-text matching, a **bridge connecting image and language**, is an important task, which generally learns a holistic cross-modal embedding to achieve a high-quality semantic alignment between the two modalities.

- ✓ The critical challenge is accurately and efficiently **learning cross-modal embeddings** and their similarities for images and texts, to achieve a high-quality semantic alignment.

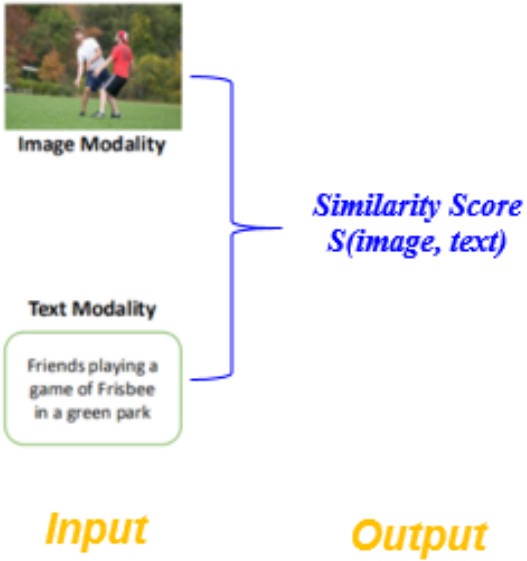


Image-text Matching

- ✓ In general, existing image-text matching methods can be classified into **two paradigms**.
- ✓ The first **embedding-based** matching separately encodes the whole images and texts into a holistic embedding space, then globally measures the semantic similarity of the two modalities.
- ✓ The second **score-based** matching applies the cross-modal interaction between visual and textual local features, then learns a cumulative similarity score.

embedding-based

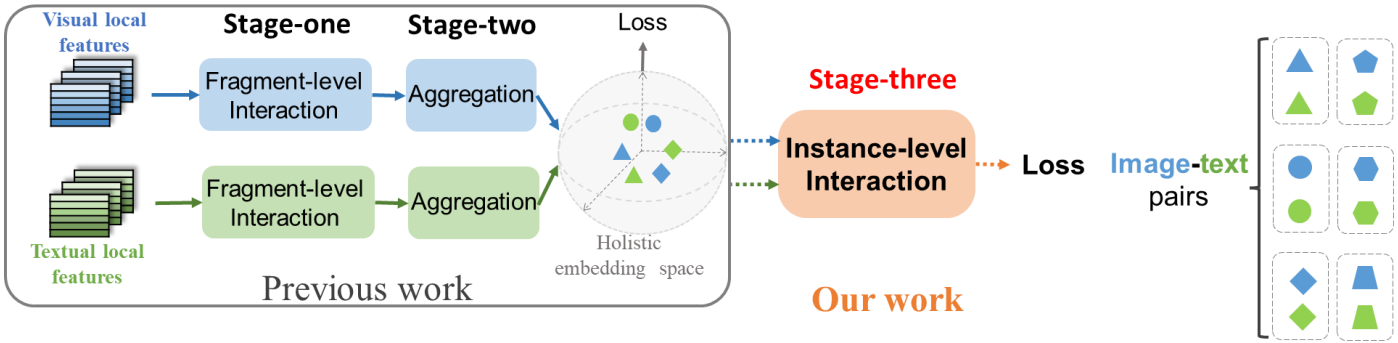
score-based

Motivation

✓ Recently, embedding-based methods have served as the mainstream solution owing to both accuracy and efficiency in image-text matching, which contains **two steps**:

(1) Capturing the intra-modal relation between visual fragments (*e.g.,* regional features) or textual fragments (*e.g.,* word features) independently, then enhancing the semantic representation of local features.

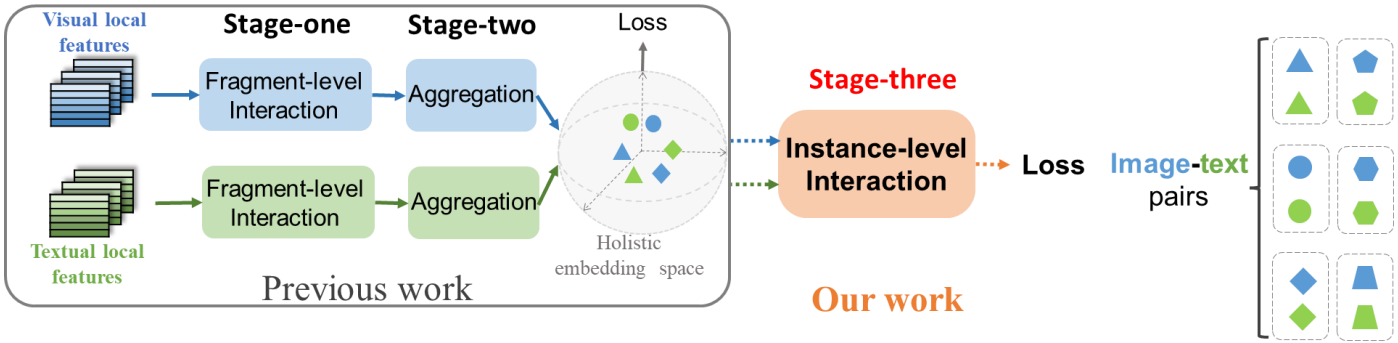
(2) Aggregating relation-enhanced local features of two modalities into the holistic embedding space.



(a) Pipeline of embedding-based method

Motivation

- ✓ Existing embedding-based methods only focus on the fragment-level relation modeling and local features interaction within one sample, *e.g.,* the region features inside one image (or the word features inside one text.
- ✓ In this way, the instance-level relation modeling and global embeddings interaction among different samples and modalities, *e.g.,* holistic embeddings of multiple images and texts, are entirely overlooked.



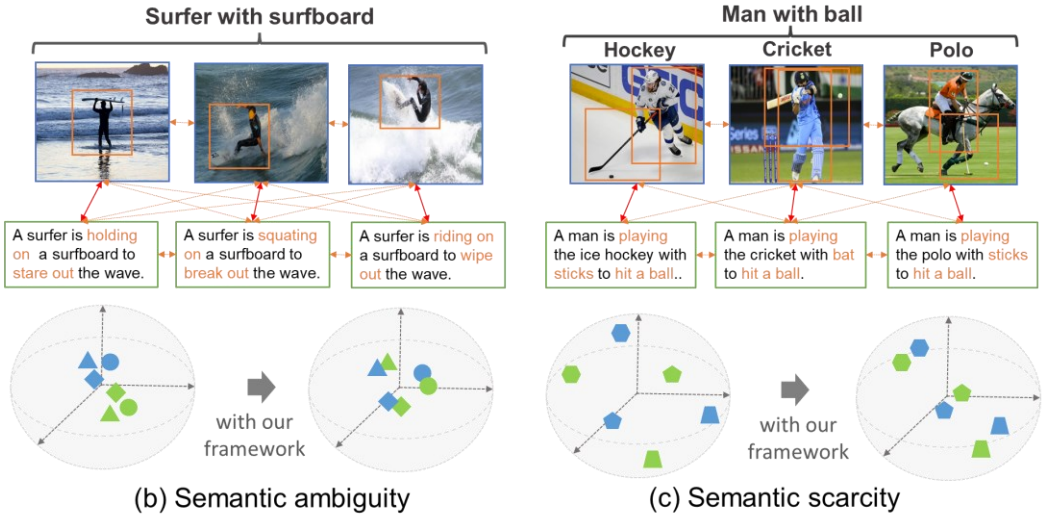
(a) Pipeline of embedding-based method

Motivation

✓ Consequently, existing embedding-based methods bring two problem.

(1) They fail to learn subtle semantic discrepancies among different samples, then can not distinguish hard negative samples with semantic ambiguities because of the heterogeneity of visual and textual semantics.

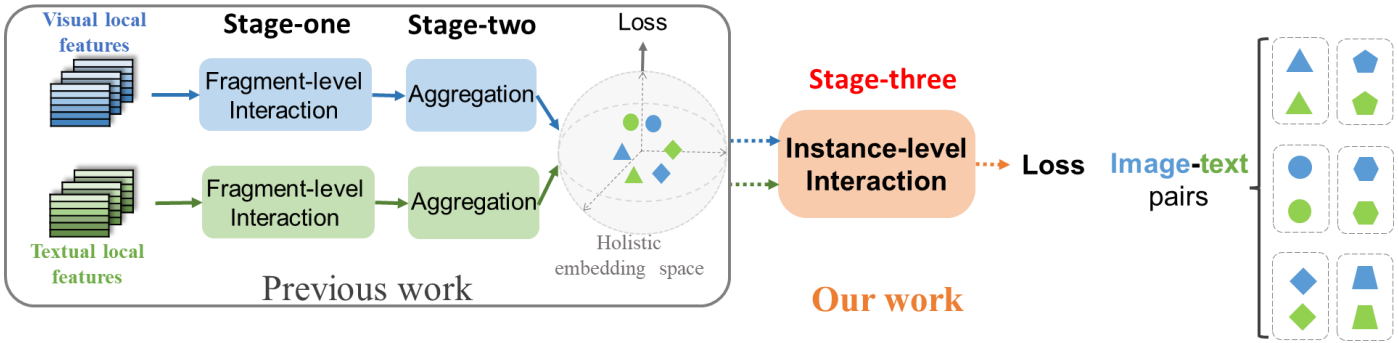
(2) They are unable to transfer shared knowledge from diverse sample, then can not effectively learn on these infrequent samples with semantic scarcities.



Contribution

- ✓ In doing so, we propose a **H**ierarchical **RE**lation **M**odeling framework (HREM) that, for the first time to our knowledge, explicitly captures both fragment-level and instance-level relations to learn holistic embeddings jointly.

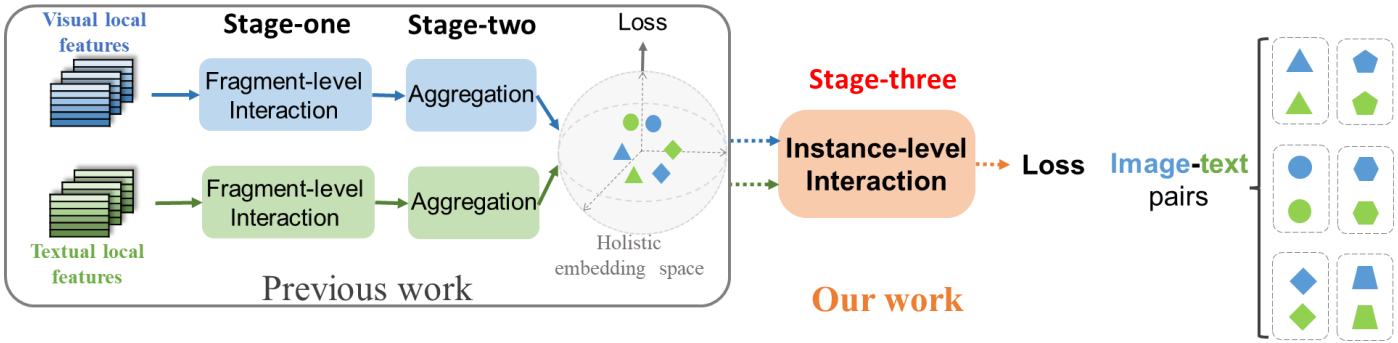
- ✓ HREM learns not only contextual semantics among intra-modal fragments to enhance local features, but also the associated semantics among inter-modal instances to distinguish hard negative samples and improve learning on infrequent samples.



(a) Pipeline of embedding-based method

Contribution

- ✓ We propose a novel **stage-three** to exactly capture the semantic relation of cross-modal samples.
- ✓ HREM only needs to capture the instance-level relation for training, then encode multi-modal embeddings independently at the inference stage, to achieve high accuracy and efficiency for image-text matching.
- ✓ Extensive experiments on Flickr30K and MS-COCO show our proposed method outperforms the state-of-the-art methods.

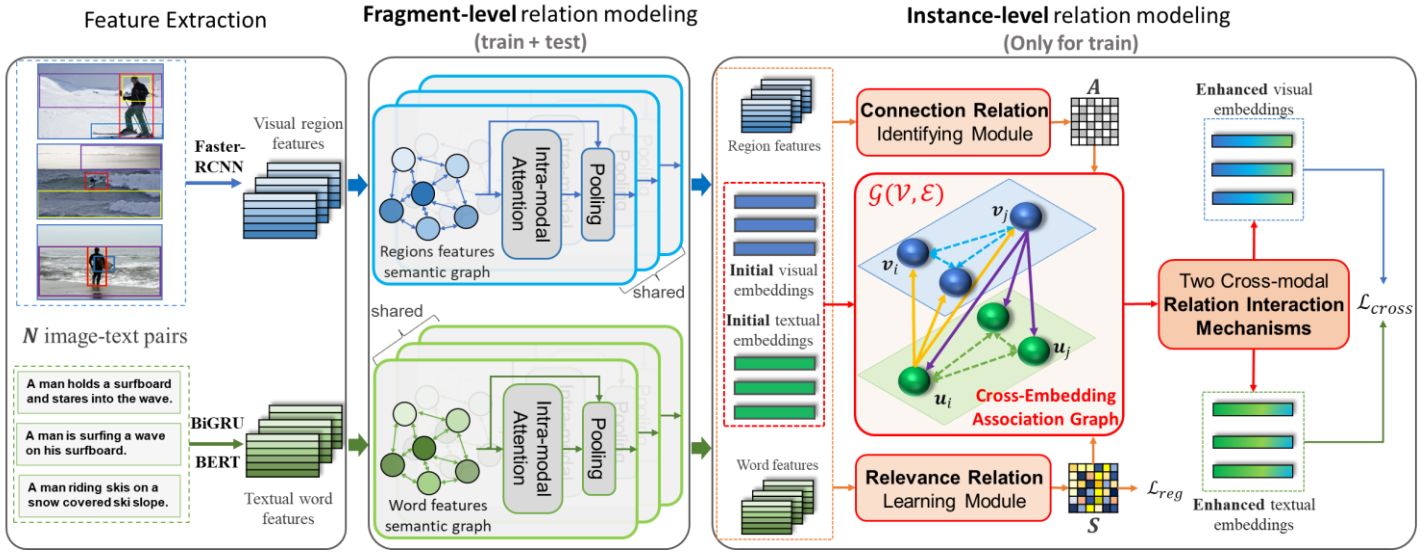


(a) Pipeline of embedding-based method

Framework

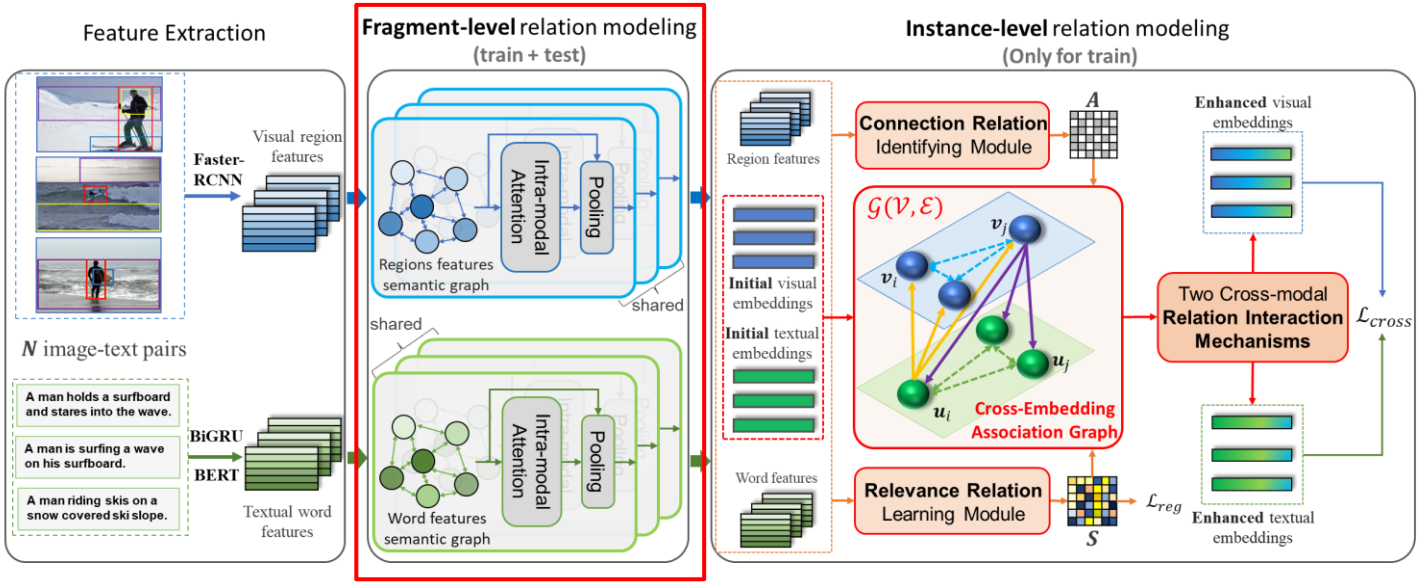
(1) We first propose a novel cross-embedding association graph, which explicitly identifies the connection relation and learns the relevance relation between batch samples with fragment-level semantic matching.

(2) Then, we propose two relation interaction mechanisms, which explore inter-modal and intra-modal relations synchronously or asynchronously with our improved attention modules to obtain enhanced embeddings



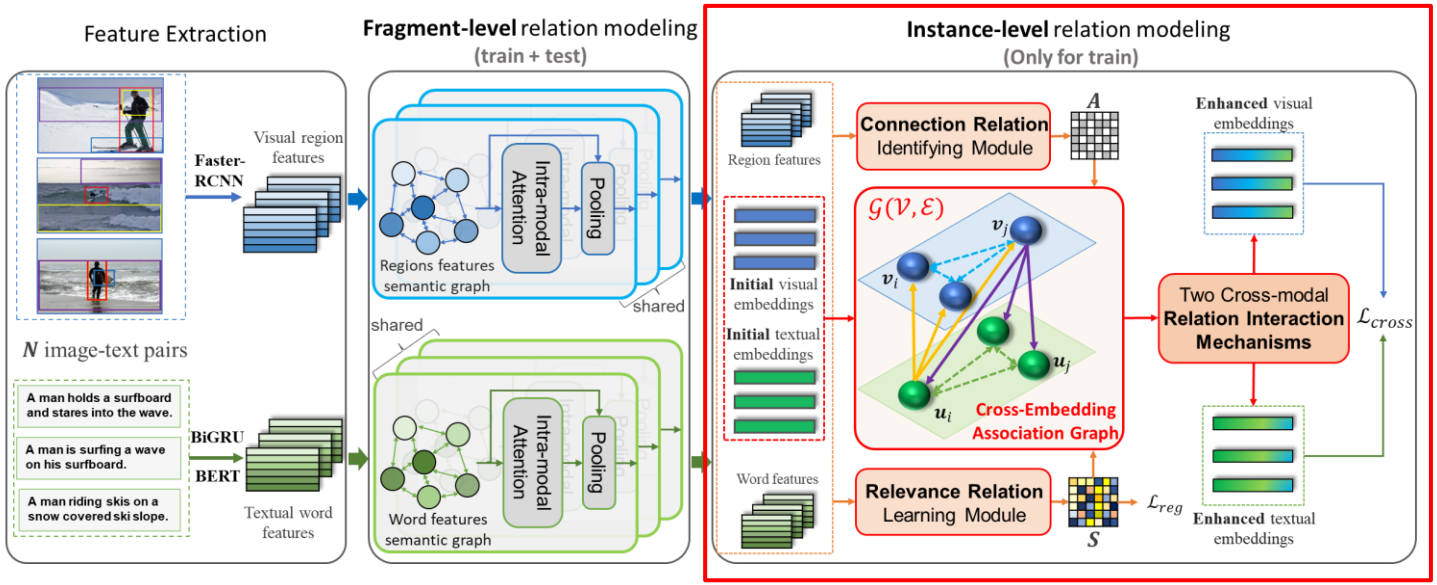
Fragment-level Relation Modeling

- ✓ Given the region and word features from the visual and textual encoder, we capture contextual information between these local features and enhance them by constructing a semantic relation graph.



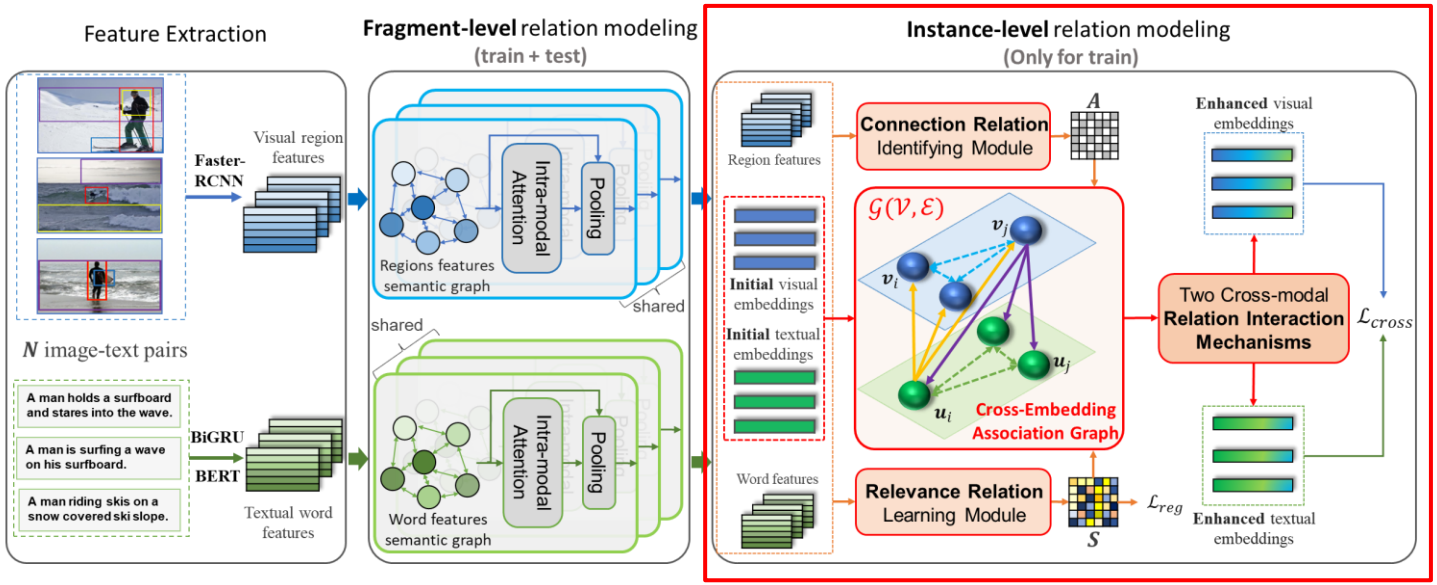
Instance-level Relation Modeling

- ✓ Given multiple image-text pairs and their embeddings, we first build a cross-embedding association graph to capture their connection relation and relevance relation, respectively.
- ✓ Then we design two relation interaction mechanisms to capture the semantic relations between multiple images and texts, where embeddings are updated by the information interaction process.



Instance-level Relation Modeling

- ✓ The instance-level relation modeling is only designed for the training stage, our framework can encode the cross-modal embeddings without sample interaction at the inference stage.



Experiments

- ✓ Extensive experiments on Flickr30K and MS-COCO show our proposed method outperforms the state-of-the-art methods.

Method	Flickr30K 1K							MS-COCO 1K						
	IMG → TEXT			TEXT → IMG			rSum	IMG → TEXT			TEXT → IMG			rSum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Region + BiGRU														
$VSRN^{*}_{2019}$ [20]	71.3	90.6	96.0	54.7	81.8	88.2	482.6	76.2	94.8	98.2	62.8	89.7	95.1	516.8
$CVSE_{2020}$ [43]	73.5	92.1	95.8	52.9	80.4	87.8	482.4	74.8	95.1	98.3	59.9	89.4	95.2	512.7
GPO_{2021} [4]	76.5	94.2	97.7	56.4	83.4	89.9	498.1	78.5	96.0	98.7	61.7	90.3	95.6	520.8
MV_{2022} [24]	79.0	94.9	97.7	59.1	84.6	90.6	505.8	78.7	95.7	98.7	62.7	90.4	95.7	521.9
HREM (Fusion)	79.5	94.3	97.4	59.3	85.1	91.2	506.8	80.0	96.0	98.7	62.7	90.1	95.4	522.8
HREM (Full)*	**81.4**	**96.5**	**98.5**	**60.9**	**85.6**	**91.3**	**514.3**	**81.2**	**96.5**	**98.9**	**63.7**	**90.7**	**96.0**	**527.1**
Region + BERT														
$CAMERA^{*}_{2020}$ [35]	78.0	95.1	97.9	60.3	85.9	91.7	508.9	77.5	96.3	98.8	63.4	90.9	95.8	522.7
$DSRAN^{*}_{2021}$ [45]	77.8	95.1	97.6	59.2	86.0	91.9	507.6	78.3	95.7	98.4	64.5	90.8	95.8	523.5
GPO_{2021} [4]	81.7	95.4	97.6	61.4	85.9	91.5	513.5	79.7	96.4	98.9	64.8	91.4	96.3	527.5
$VSRN{+}{+}_{2022}$ [21]	79.2	94.6	97.5	60.6	85.6	91.4	508.9	77.9	96.0	98.5	64.1	91.0	96.1	523.6
HREM (Fusion)	83.3	96.0	98.1	63.5	87.1	92.4	520.4	81.1	96.6	98.9	66.1	91.6	96.5	530.7
HREM (Full)*	**84.0**	**96.1**	**98.6**	**64.4**	**88.0**	**93.1**	**524.2**	**82.9**	**96.9**	**99.0**	**67.1**	**92.0**	**96.6**	**534.6**

Experiments

- ✓ Extensive experiments on Flickr30K and MS-COCO show our proposed method outperforms the state-of-the-art methods.

Type	Method	MS-COCO 5K						
		IMG → TEXT			TEXT → IMG			rSum
		R@1	R@5	R@10	R@1	R@5	R@10	
Region + BiGRU								
S	IMRAM$^{*}_{2020}$ [3]	53.7	83.2	91.0	39.7	69.1	79.8	416.5
	UARDA$^{*}_{2022}$ [50]	56.2	83.8	91.3	40.6	69.5	80.9	422.3
	NAAF$^{*}_{2022}$ [51]	58.9	85.2	92.0	**42.5**	70.9	81.4	430.9
E	GPO$_{2021}$ [4]	56.6	83.6	91.4	39.3	69.9	81.1	421.9
	CGMN$_{2022}$ [5]	53.4	81.3	89.6	41.2	71.9	82.4	419.8
	MV$_{2022}$ [24]	56.7	84.1	91.4	40.3	70.6	81.6	424.6
	HREM (Standalone)	58.4	85.5	92.4	39.8	70.5	81.0	427.6
	HREM (Fusion)	58.9	85.3	92.1	40.0	70.6	81.2	428.1
	HREM (Full)*	**60.6**	**86.4**	**92.5**	41.3	**71.9**	**82.4**	**435.1**
Region + BERT								
S	SSAMT$_{2021}$ [11]	57.7	84.2	90.8	40.8	70.5	80.5	424.5
	DIME$^{*}_{2021}$ [36]	59.3	85.4	91.9	43.1	73.0	83.1	435.8
	DCPA$_{2022}$ [40]	53.5	82.4	90.2	40.4	71.0	82.0	419.5
E	DSRAN$^{*}_{2021}$ [45]	55.3	83.5	90.9	41.7	72.7	82.8	426.9
	GPO$_{2021}$ [4]	58.3	85.3	92.3	42.4	72.7	83.2	434.3
	VSRN++$_{2022}$ [21]	54.7	82.9	90.9	42.0	72.2	82.7	425.4
	HREM (Standalone)	61.8	87.0	93.2	44.0	73.7	83.4	443.1
	HREM (Fusion)	62.3	87.6	93.4	43.9	73.6	83.3	444.1
	HREM (Full)*	**64.0**	**88.5**	**93.7**	**45.4**	**75.1**	**84.3**	**450.9**

Ablation Study

✓ The ablation study show our hierarchical relation modeling methods are meaningful.

(a) The ablation study of hierarchical relation modeling on Flickr30K

Fragment-level		Instance-level		IMG → TEXT		TEXT → IMG	
Visual	Textual	Intra-modal	Inter-modal	R@1	R@5	R@1	T R@5
		✓	✓	81.5	94.9	62.3	86.2
	✓	✓	✓	81.8	95.1	62.4	86.2
✓		✓	✓	83.1	95.9	63.2	87.0
✓	✓			80.1	94.8	60.9	85.3
✓	✓	✓		80.5	94.7	61.2	85.2
✓	✓		✓	82.2	95.5	62.6	86.4
✓	✓	✓	✓	83.3	96.0	63.5	87.1

(b) The ablation study of instance-level relation modeling on Flickr30K

Methods	IMG → TEXT		TEXT → IMG	
	R@1	R@5	R@1	T R@5
w/o connection matrix $\mathbf{A}$	81.4	95.1	61.5	86.3
w/o relevance matrix $\mathbf{S}$	81.7	95.3	61.9	86.5
w/o consistency $\mathcal{L}_{cross}$	81.6	95.6	61.8	86.7
w/o regularization $\mathcal{L}_{reg}$	82.8	95.8	62.9	86.9
w/o neighbor batch sampling	82.8	95.9	63.1	87.0
HREM	83.3	96.0	63.5	87.1

Performance vs Run-time

- ✓ Although our method belongs to the embedding-based image-text matching methods，achieves both high accuracy and efficiency on cross-modal retrieval.