Improving Cross-Modal Retrieval with Set of Diverse Embeddings

Dongwon Kim Namyup Kim Suha Kwak
POSTECH (Pohang University of Science and Technology)

TL;DR: "We propose an *efficient* & *effective* cross-modal retrieval method that represents a sample with *diverse* embeddings."

Motivation

Cross-modal retrieval

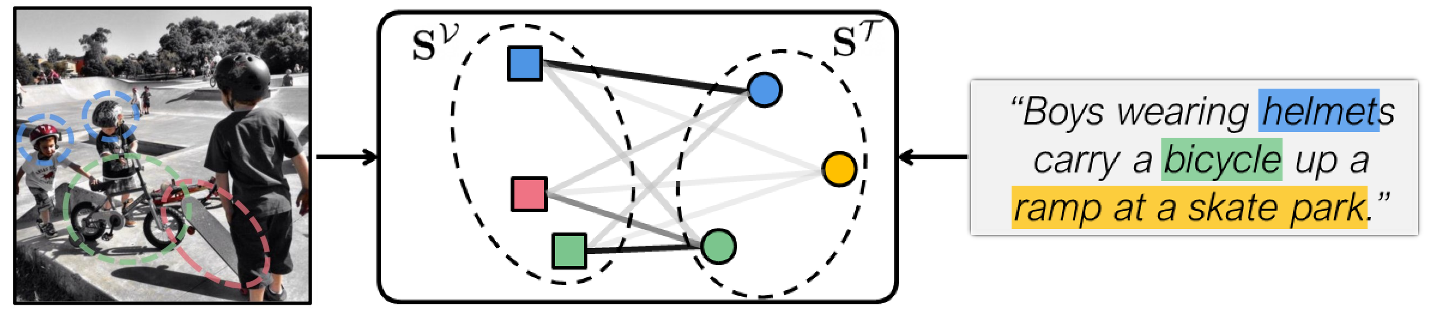
- Searching for data when the query and database have different modalities (image and text).

Ambiguity problem

- Even a single image often contains various contexts.
- Visual manifestations of a caption vary significantly.

Previous work: Set-based embedding

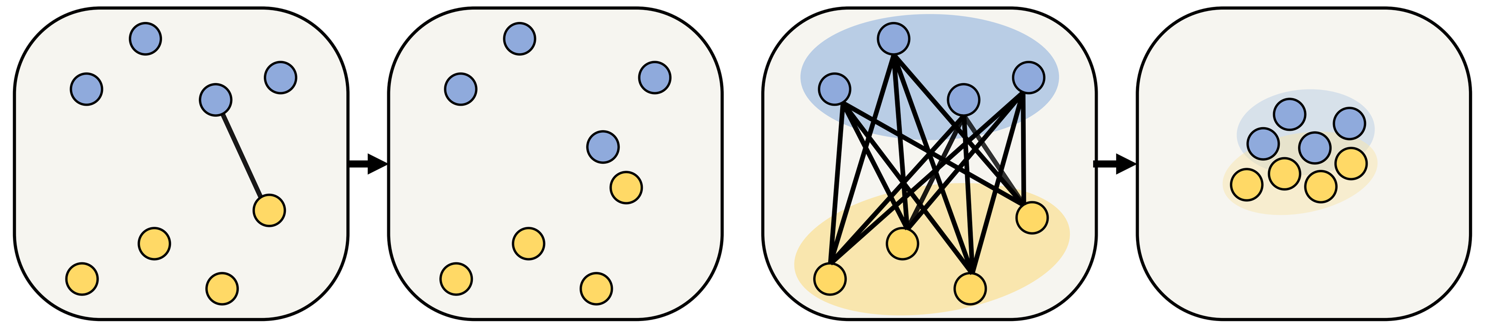
- Represent the data with the ***set of embedding vectors*** *(embedding set)* [1,2].
- Ambiguity of the data is addressed by elements of the embedding set, which represent diverse semantics.



Embedding space

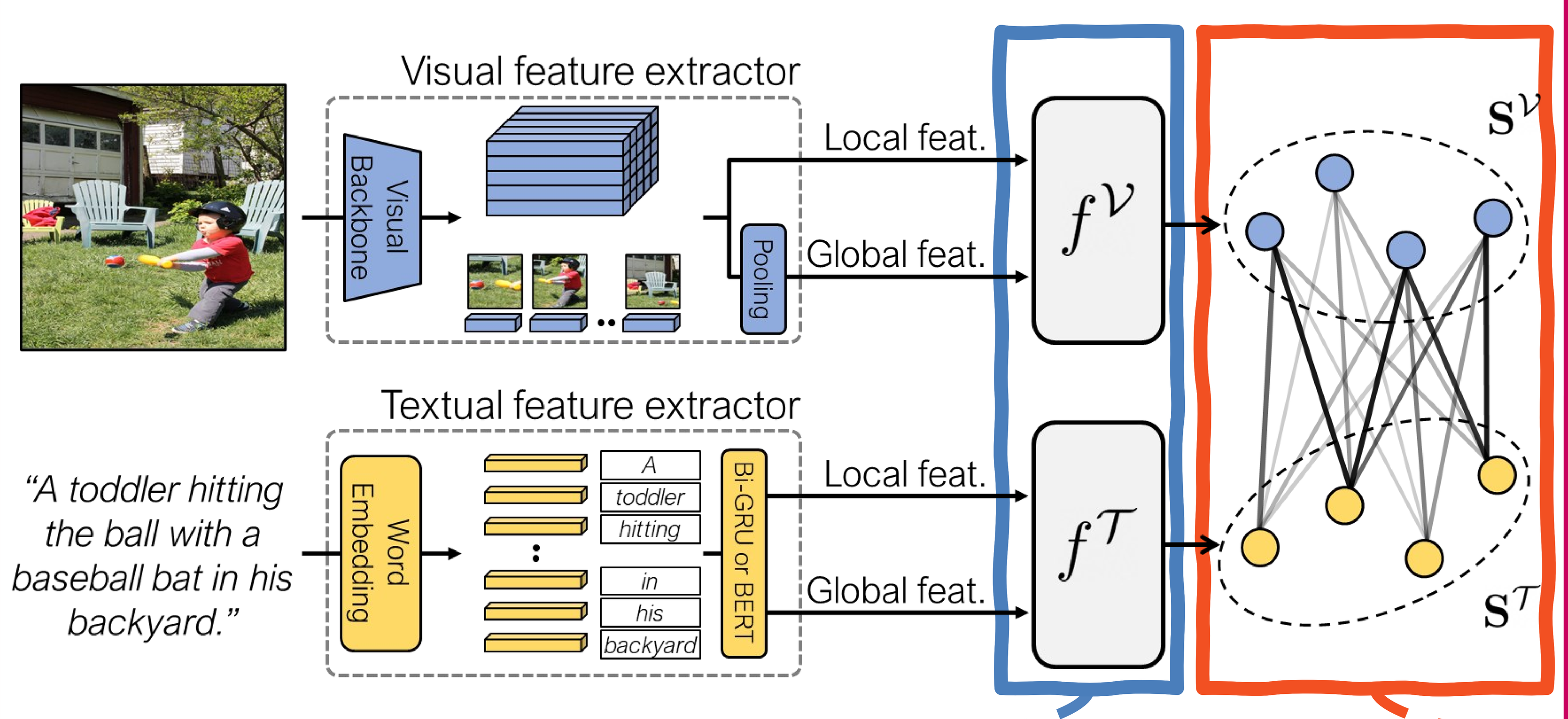
Drawbacks of previous set-based embedding

- ***Sparse supervision***→ An embedding set most of whose elements remain untrained.
- ***Set collapsing***→ An embedding set with a small variance which does not encode sufficient ambiguity.



Sparse supervision **Set collapsing**

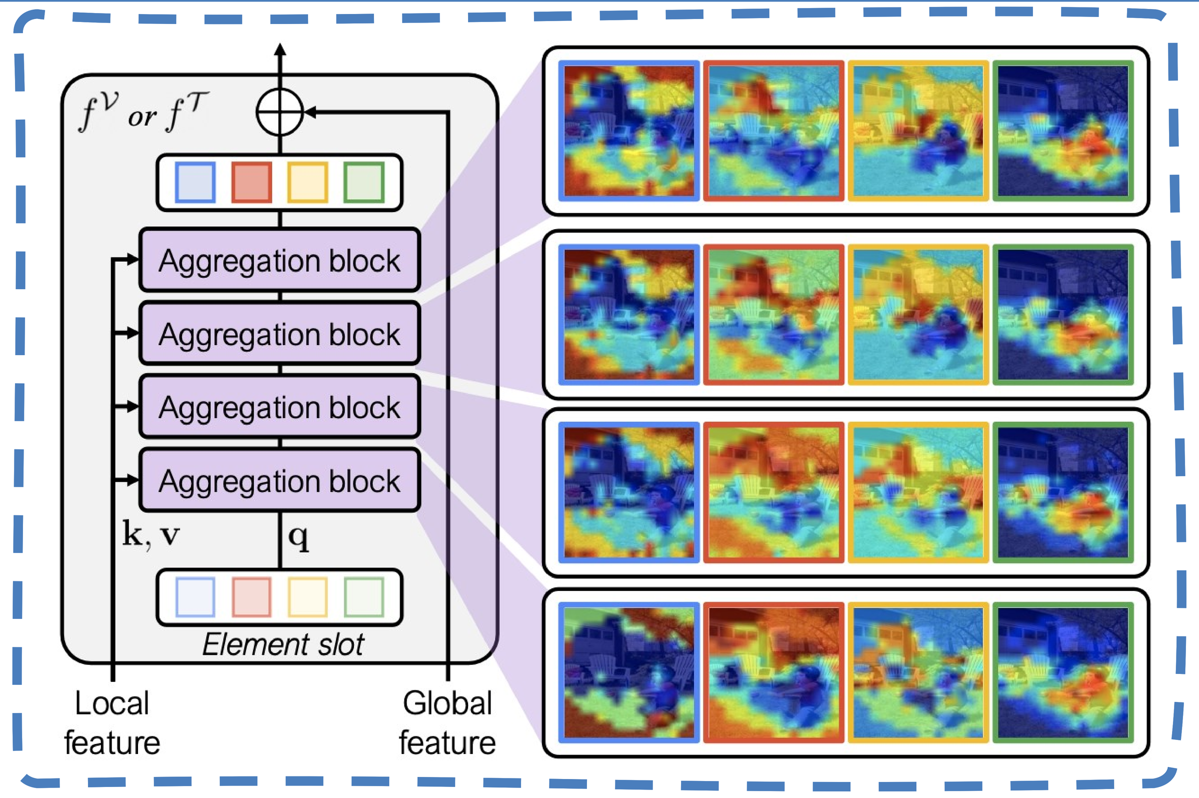
Our solutions



- ***Smooth-Chamfer similarity:*** Similarity function between sets that provides *dense supervision without collapsing.*
- ***Set-prediction module:*** The module captures *diverse semantic ambiguity* of input, motivated by slot-attn [3].

Set-prediction module

- Element slots ***compete*** for aggregating input, progressively transformed into embedding set.
- ***Competition between slots*** makes element encode substantially different semantics.



$$\text{attn} = \text{softmax}\left(\frac{1}{\sqrt{D}}k(\text{inputs})\cdot q(\text{slots})^T, \boxed{\text{axis='slots'}}\right)$$

Slot-attn [3] based attention scheme (**Ours**)

$$\text{attn} = \text{softmax}\left(\frac{1}{\sqrt{D}}k(\text{inputs})\cdot q(\text{slots})^T, \boxed{\text{axis='inputs'}}\right)$$

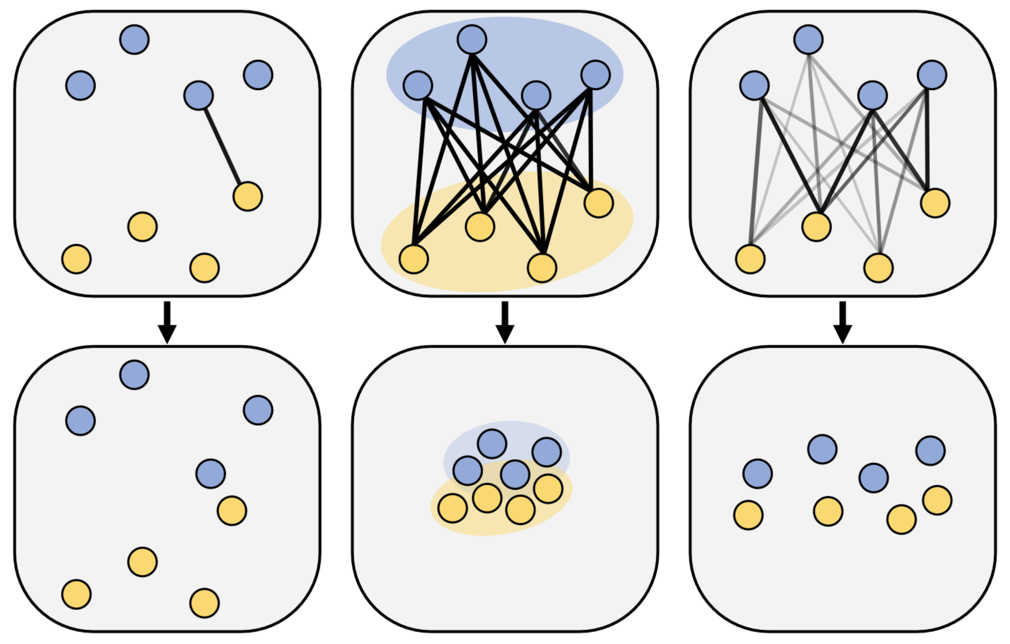
Conventional transformer attention scheme

Smooth-Chamfer similarity

Proposed SC similarity associates

- every possible pair
  → *Resolves **sparse supervision***
- with different degree of weights.
  → *Resolves **set collapsing***

$$s_{\text{SC}}(\mathbf{S}_1, \mathbf{S}_2) = \frac{1}{2\alpha|\mathbf{S}_1|}\sum_{x\in\mathbf{S}_1}\text{LSE}_{y\in\mathbf{S}_2}(\alpha c(x,y)) + \frac{1}{2\alpha|\mathbf{S}_2|}\sum_{y\in\mathbf{S}_2}\text{LSE}_{x\in\mathbf{S}_1}(\alpha c(x,y))$$



(a) MIL similarity (b) Match probability (c) Smooth-Chamfer similarity (**Ours**)

Experiments

Achieved SOTA on COCO, Flickr30K, CxC, and ECCV-caption

Method	CA	1K Test Images: Image-to-Text R@1	R@5	R@10	Text-to-Image R@1	R@5	R@10	RSUM	5K Test Images: Image-to-Text R@1	R@5	R@10	Text-to-Image R@1	R@5	R@10	RSUM
Faster R-CNN + Bi-GRU															
SCAN† [30]	✓	72.7	94.8	98.4	58.8	88.4	94.8	507.9	50.4	82.2	90.0	38.6	69.3	80.4	410.9
VSRN† [31]	✗	76.2	94.8	98.2	62.8	89.7	95.1	516.8	53.0	81.1	89.4	40.5	70.6	81.1	415.7
CAAN [53]	✓	75.5	95.4	98.5	61.3	89.7	95.2	515.6	52.5	83.3	90.9	41.2	70.3	82.9	421.1
IMRAM† [6]	✓	76.7	95.6	98.5	61.7	89.1	95.0	516.6	53.7	83.2	91.0	39.7	69.1	79.8	416.5
SGRAF† [14]	✓	79.6	96.2	98.5	63.2	90.7	96.1	524.3	57.8	-	91.6	41.9	-	81.3	-
VSE_∞ [27]	✗	78.5	96.0	98.7	61.7	90.3	95.6	520.8	56.6	83.6	91.4	39.3	69.9	81.1	421.9
NAAF† [52]	✓	80.5	96.5	98.8	64.1	90.7	96.5	527.2	58.9	85.2	92.0	42.5	70.9	81.4	430.9
Ours	✗	79.8	96.2	98.6	63.6	90.7	95.7	524.6	58.8	84.9	91.5	41.1	72.0	82.4	430.7
Ours†	✗	80.6	96.3	98.8	64.7	91.4	96.2	**528.0**	60.4	86.2	92.4	42.6	73.1	83.1	**437.8**
ResNeXt-101 + BERT															
VSE_∞ [27]	✗	84.5	98.1	99.4	72.0	93.9	97.5	545.4	66.4	89.3	94.6	51.6	79.3	87.6	468.9
VSE_∞† [27]	✗	85.6	98.0	99.4	73.1	94.3	97.7	548.1	68.1	90.2	95.2	52.7	80.2	88.3	474.8
Ours	✗	86.3	97.8	99.4	72.4	94.0	97.6	547.5	69.1	90.7	95.6	52.1	79.6	87.8	474.9
Ours†	✗	86.6	98.2	99.4	73.4	94.5	97.8	**549.9**	71.0	91.8	96.3	53.4	80.9	88.6	**482.0**

Embedding set elements & their nearest caption

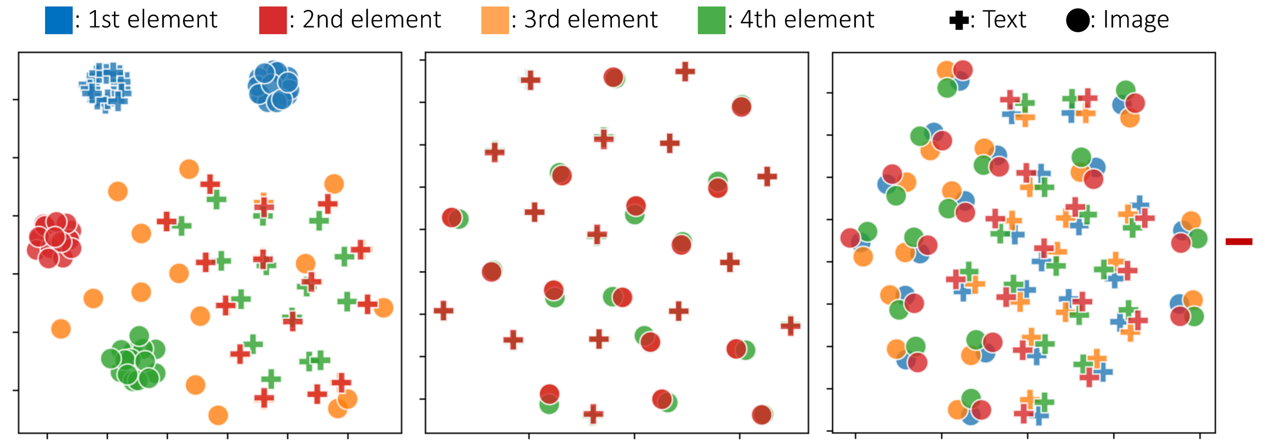
Ablation studies: similarity and model

Similarity	Arch.	RSUM
MIL	Ours	491.7
MP	Ours	490.5
Ours (Chamfer)	Ours	499.6
Ours (S-Chamfer)	PIE-Net	483.3
Ours (S-Chamfer)	Ours	**500.8**

Setting	log(Var.)	RSUM
PIE-Net	-7.35	483.3
Ours \w MP	-5.27	490.5
Transformer	-2.27	496.1
Ours	-2.13	**500.8**

*Our method successfully resolves **sparse supervision** & **set collapsing** issues.*

Embedding space visualization



(a) MIL (b) Match Probability (c) Ours

[1] Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval, CVPR, 2019.
[2] Probabilistic Embeddings for Cross-Modal Retrieval, CVPR, 2021.
[3] Object-centric learning with slot attention, NeurIPS, 2020.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA