Improving Cross-modal Retrieval with Set of Diverse Embeddings

Dongwon Kim
kdwon@postech.ac.kr

Namyup Kim
namyup@postech.ac.kr

Suha Kwak
suha.kwak@postech.ac.kr
Cross-modal Retrieval

Text-to-image

Children riding bikes and skateboards

Image-to-text

Boys wearing helmets carry a bike up a ramp at a skate park.
Small children stand near bicycles at a skate park.
A group of young children riding bikes and skateboards.
Semantic Ambiguity

An image or a sentence often illustrates multiple entities and their relations.

“Boys wearing helmets carry a bicycle up a ramp at a skate park.”

“Small children stand near bicycles at a skate park.”

“A group of young children riding bikes and skateboards.”
It is impractical to manually annotate such entities and their correspondences.

“Boys wearing helmets carry a bicycle up a ramp at a skate park.”

“Small children stand near bicycles at a skate park.”

“A group of young children riding bikes and skateboards.”
Embedding Network Architectures

**Single Cross-attention Encoder**

Similarity: \( g(x, y) \)

Cross-attention encoder \( g \)

- \( x \): Small children stand near bicycles at a skate park.
- \( y \): Image encoder + Text encoder

\[ \text{Similarity: } s \left( f^V(x), f^T(y) \right) \]

**Image Encoder + Text Encoder**

- \( x \): Image encoder \( f^V \)
- \( y \): Text encoder \( f^T \)

- \( x \): Small children stand near bicycles at a skate park.
- \( y \): Image encoder + Text encoder
Embedding Network Architectures

**Single Cross-attention Encoder**

Similarity: \( g(x, y) \)

(+) Boosting performance by fine-grained image-text interaction

(−) Impractical for large-scale image retrieval due to the prohibitively heavy computation at inference

**Image Encoder + Text Encoder**

Similarity: \( s(f^V(x), f^T(y)) \)

(+) Appropriate for large-scale image retrieval thanks to the simple and efficient similarity computation

(−) Limited performance due to the lack of image-text interaction
Our Approach

① Separate encoders for efficient retrieval

② Embedding set representation + set similarity metric for resolving the ambiguity issue

A toddler hitting the ball with a baseball bat in his backyard.
Contribution

• A new set-based embedding architecture
  • Set-prediction modules based on slot attention

• A new set similarity metric
  • Smooth-Chamfer similarity

• Outstanding performance
  • State of the art in most settings on four public benchmarks
  • Leading to substantially less latency than cross-attention models
Proposed Architecture

A toddler hitting the ball with a baseball bat in his backyard.
Proposed Architecture: Set Prediction Modules

The element slots\(^1\) compete with each other to aggregate input features and thus reveal diverse contexts.

[1] Locatello et al., Object-centric Learning with Slot Attention, NeurIPS 2020.
Proposed Architecture: Set Prediction Modules

Local features $\psi \rightarrow$ (Key, Value) pairs: $k, v \in \mathbb{R}^{N \times D_h}$
Element slots $E_{t-1} \rightarrow$ Queries: $q \in \mathbb{R}^{K \times D_h}$

**Computing an attention map**

$$A_{n,k} = \frac{\exp M_{n,k}}{\sum_{i=1}^K \exp M_{n,i}}, \text{ where } M = \frac{kq^T}{\sqrt{D_h}}$$

Normalization over the slots\(^{[1]}\)

**Updating the element slots**

$$E_t = \text{MLP}(\bar{E}^t) + \bar{E}^t, \text{ where } \bar{E}^t = \hat{A}^T v W_0 + E_{t-1}$$

and

$$\hat{A}_{n,k} = \frac{A_{n,k}}{\sum_{i=1}^N A_{n,k}}$$

---

\(^{[1]}\) Locatello et al., Object-centric Learning with Slot Attention, NeurIPS 2020.
Proposed Architecture: Set Prediction Modules

Adding the global feature to each element

\[ S = \text{LN}(E) + [\text{LN}(\phi), \ldots, \text{LN}(\phi)] \in \mathbb{R}^{K \times D} \]

- Embedding the global context in every element of the set
- Particularly useful when treating samples with little ambiguity
Set Similarity Metric: Smooth-Chamfer Similarity

\[ s(S^\nu, S^\tau) = \frac{1}{2\alpha |S^\nu|} \sum_{e \in S^\nu} \sum_{e' \in S^\tau} \text{LSE} (\alpha \cos(e, e')) + \frac{1}{2\alpha |S^\tau|} \sum_{e' \in S^\tau} \sum_{e \in S^\nu} \text{LSE} (\alpha \cos(e, e')) \]

\[
\log \left( \sum_{y \in S_2} \exp[\alpha \cos(x, y)] \right) \quad \log \left( \sum_{x \in S_1} \exp[\alpha \cos(x, y)] \right)
\]
Set Similarity Metric: Smooth-Chamfer Similarity

\[ s(S^V, S^T) = \frac{1}{2\alpha |S^V|} \sum_{e \in S^V, e' \in S^T} \text{LSE} (\alpha \cos(e, e')) + \frac{1}{2\alpha |S^T|} \sum_{e' \in S^T, e \in S^V} \text{LSE} (\alpha \cos(e, e')) \]
Training Objective

\[ \mathcal{L}\left(\{\mathbf{s}_i^v, \mathbf{s}_i^T\}_{i=1}^N\right) = \mathcal{L}_{\text{tri}}\left(\{\mathbf{s}_i^v, \mathbf{s}_i^T\}_{i=1}^N\right) + \mathcal{L}_{\text{mmd}}\left(\{\mathbf{s}_i^v\}_{i=1}^N, \{\mathbf{s}_i^T\}_{i=1}^N\right) + \mathcal{R}_{\text{div}} \]

Metric learning

A boy hitting the ball with a baseball bat in his backyard.

Small children stand near bicycles at a skate park.

(\mathbf{x}_i, \mathbf{y}_i)

(\mathbf{x}_j, \mathbf{y}_j)
Training Objective

\[ \mathcal{L} \left( \{ \mathbf{S}_i^\nu, \mathbf{S}_i^T \}_{i=1}^N \right) = \mathcal{L}_{\text{tri}} \left( \{ \mathbf{S}_i^\nu, \mathbf{S}_i^T \}_{i=1}^N \right) + \mathcal{L}_{\text{mmd}} \left( \{ \mathbf{S}_i^\nu \}_{i=1}^N, \{ \mathbf{S}_i^T \}_{i=1}^N \right) + \mathcal{R}_{\text{div}} \]

Closing the modality gap
Training Objective

\[ \mathcal{L}\left(\{s_i^y, s_i^T\}_{i=1}^N\right) = \mathcal{L}_{\text{tri}}\left(\{s_i^y, s_i^T\}_{i=1}^N\right) + \mathcal{L}_{\text{mmd}}\left(\{s_i^y\}_{i=1}^N, \{s_i^T\}_{i=1}^N\right) + R_{\text{div}} \]

Enhancing within-set diversity

A boy hitting the ball with a baseball bat in his backyard.

Small children stand near bicycles at a skate park.
Training Objective

\[ \mathcal{L} \left( \{ \mathbf{s}_i^y, \mathbf{s}_i^T \}_{i=1}^N \right) = \mathcal{L}_{\text{tri}} \left( \{ \mathbf{s}_i^y, \mathbf{s}_i^T \}_{i=1}^N \right) + \mathcal{L}_{\text{mmd}} \left( \{ \mathbf{s}_i^y \}_{i=1}^N, \{ \mathbf{s}_i^T \}_{i=1}^N \right) + \mathcal{R}_{\text{div}} \]

**Triplet rank loss with hard negative mining**

\[ \mathcal{L}_{\text{tri}} \left( \{ \mathbf{s}_i^y, \mathbf{s}_i^T \}_{i=1}^N \right) = \sum_{i=1}^{N} \max \left[ \delta + s(\mathbf{s}_i^y, \mathbf{s}_j^T) - s(\mathbf{s}_i^y, \mathbf{s}_i^T) \right]_+ + \sum_{i=1}^{N} \max \left[ \delta + s(\mathbf{s}_i^T, \mathbf{s}_j^y) - s(\mathbf{s}_i^T, \mathbf{s}_i^y) \right]_+ \]

**Maximum mean discrepancy\(^{[2]}\) loss**

\[ \mathcal{L}_{\text{mmd}} \left( \{ \mathbf{s}_i^y \}_{i=1}^N, \{ \mathbf{s}_i^T \}_{i=1}^N \right) = \text{MMD} \left( \{ \mathbf{s}_i^y \}_{i=1}^N, \{ \mathbf{s}_i^T \}_{i=1}^N \right) \]

**Diversity regularizer**

\[ \mathcal{R}_{\text{div}} = \sum_{e, e' \in \mathcal{E}} \exp(-2\|e - e'\|_2^2) \]

[2] Gretton et al., A Kernel Two-sample Test, JMLR 2012.
Experiments

• Datasets
  • COCO\textsuperscript{3}, Flickr30K\textsuperscript{4}, ECCV Caption\textsuperscript{5}, CrissCrossed Caption (CxC)\textsuperscript{6}

• Evaluation metrics
  • \textbf{Recall}@\textit{k}: Percentage of the queries that have matching samples among top-\textit{k} retrieval results
  • \textbf{RSUM}: Sum of Recall@\textit{k} at \textit{k} \in \{1, 5, 10\} in both image-to-text and text-to-image settings

• 4 agg. blocks and 4 element slots for each set-prediction module

\textsuperscript{3} Lin \textit{et al.}, Microsoft COCO: Common Objects in Context, ECCV 2014.
\textsuperscript{4} Plummer \textit{et al.}, Flickr30k Entities: Collecting Region-to-phrase Correspondences for Richer Image-to-sentence Models, ICCV 2015.
\textsuperscript{5} Chun \textit{et al.}, ECCV Caption, Correcting False Negatives by Collecting Machine-and-human-verified Image-Caption Associations for MS-COCO, ECCV 2022.
\textsuperscript{6} Parekh \textit{et al.}, Crisscrossed Captions: Extended Intra-modal and Inter-modal Semantic Similarity Judgments for MS-COCO, EACL 2020.
## Experiments: Performance on COCO

| Method          | CA | Image-to-Text |             | Text-to-Image | RSUM | Image-to-Text |             | Text-to-Image | 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 | R@1 | R@5 | R@10 |
| **ResNet-152 + Bi-GRU** |    |      |     |      |      |     |      |      |     |      |      |     |      |      |     |      |      |
| VSE++           | ✗  | 64.6 | 90.0 | 95.7 | 52.0 | 84.3 | 92.0 | 478.6 | 41.3 | 71.1 | 81.2 | 30.3 | 59.4 | 72.4 | 355.7 |
| PVSE            | ✗  | 69.2 | 91.6 | 96.6 | 55.2 | 86.5 | 93.7 | 492.8 | 45.2 | 74.3 | 84.5 | 32.4 | 63.0 | 75.0 | 374.4 |
| PCME            | ✗  | 68.8 | -     | -    | 54.6 | -     | - | - | 44.2 | - | - | 31.9 | - | - | - | - |
| Ours            | ✗  | 70.3 | 91.5 | 96.3 | 56.0 | 85.8 | 93.3 | **493.2** | 47.2 | 74.8 | 84.1 | 33.8 | 63.1 | 74.7 | **377.7** |
| **Faster R-CNN + Bi-GRU** |    |      |     |      |      |     |      |      |      |     |      |      |     |      |      |     |      |      |
| SCAN†           | ✓   | 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†           | ✓   | 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            | ✓   | 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†          | ✓   | 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†          | ✓   | 79.6 | 96.2 | 98.5 | 63.2 | 90.7 | 96.1 | 524.3 | 57.8 | - | 91.6 | 41.9 | - | 81.3 | - |
| VSE∞            | ✗  | 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†           | ✓   | 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** |
| **ResNetXt-101 + BERT** |    |      |     |      |      |     |      |      |      |     |      |      |     |      |      |     |      |      |
| VSE∞            | ✗  | 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∞†          | ✗  | 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** |
## Experiments: Performance on Flickr30K

| Method          | CA | Image-to-text | Text-to-image | RSUM |
|-----------------|----|---------------|--------------|------|
| ResNet-152 + Bi-GRU |    |              |              |      |
| VSE++           | ✗  | 52.9         | 80.5         | 87.2 | 39.6  | 70.1 | 79.5 | 409.8 |
| PVSE*           | ✗  | 59.1         | 84.5         | 91.0 | 43.4  | 73.1 | 81.5 | 432.6 |
| PCME*           | ✗  | 58.5         | 81.4         | 89.3 | 44.3  | 72.7 | 81.9 | 428.1 |
| Ours            | ✗  | 61.8         | 85.5         | 91.1 | 46.1  | 74.8 | 83.3 | **442.6** |

Faster R-CNN + Bi-GRU

| Method          | CA | Image-to-text | Text-to-image | RSUM |
|-----------------|----|---------------|--------------|------|
| SCAN†           | ✓  | 67.4         | 90.3         | 95.8 | 48.6  | 77.7 | 85.2 | 465.0 |
| VSRN†           | ✗  | 71.3         | 90.6         | 96.0 | 54.7  | 81.8 | 88.2 | 482.6 |
| CAAN            | ✓  | 70.1         | 91.6         | 97.2 | 52.8  | 79.0 | 87.9 | 478.6 |
| IMRAM†          | ✓  | 74.1         | 93.0         | 96.6 | 53.9  | 79.4 | 87.2 | 484.2 |
| SGRAF†          | ✓  | 77.8         | 94.1         | 97.4 | 58.5  | 83.0 | 88.8 | 499.6 |
| VSE∞            | ✗  | 76.5         | 94.2         | 97.7 | 56.4  | 83.4 | 89.9 | 498.1 |
| NAAF†           | ✓  | 81.9         | 96.1         | 98.3 | 61.0  | 85.3 | 90.6 | **513.2** |
| Ours            | ✗  | 77.8         | 94.0         | 97.5 | 57.5  | 84.0 | 90.0 | 500.8 |
| Ours†           | ✗  | 80.9         | 94.7         | 97.6 | 59.4  | 85.6 | 91.1 | 509.3 |

ResNeXt-101 + BERT

| Method          | CA | Image-to-text | Text-to-image | RSUM |
|-----------------|----|---------------|--------------|------|
| VSE∞            | ✗  | 88.4         | 98.3         | 99.5 | 74.2  | 93.7 | 96.8 | 550.9 |
| VSE∞†           | ✗  | 88.7         | 98.9         | 99.8 | 76.1  | 94.5 | 97.1 | 555.1 |
| Ours            | ✗  | 88.8         | 98.5         | 99.6 | 74.3  | 94.0 | 96.7 | 551.9 |
| Ours†           | ✗  | 90.6         | 99.0         | 99.6 | 75.9  | 94.7 | 97.3 | **557.1** |
Experiments: Performance on Flickr30K

**Computation Complexity**

| Method       | \(\log(\text{FLOPS})\) |
|--------------|--------------------------|
| VSE\(_\infty^7\) | × 16                     |
| Ours         | × 1,280                  |
| NAAF\[^{8}\] (SCAN\[^{9}\]) |                          |

**Latency in inference**

| Method       | Latency (ms) |
|--------------|--------------|
| VSE\(_\infty^7\) | 159          |
| Ours         | 168          |
| NAAF\[^{8}\] (SCAN\[^{9}\]) | 198,121      |

[7] Jiacheng et al., Learning the Best Pooling Strategy for Visual Semantic Embedding, CVPR 2021.
[8] Zhang et al., Negative-aware Attention Framework for Image-text Matching, CVPR 2022.
[9] Lee et al., Stacked Cross Attention for Image-text Matching, ECCV 2018.
### Experiments: Performance on ECCV Caption and CxC

|       | Image-to-text |       | Text-to-image |       |
|-------|---------------|-------|---------------|-------|
|       | ECCV Caption  | CxC   | ECCV Caption  | CxC   |
|       | mAP@R R-P R@1 | R@1   | mAP@R R-P R@1 | R@1   |
| VSRN  | 30.8 42.9 73.8 | 55.1  | 53.8 60.8 89.2 | 42.6  |
| VSE∞  | 34.8 45.4 81.1 | 67.9  | 50.0 57.5 91.8 | 53.7  |
| Ours  | **36.0 46.4 84.7 72.3** |       | **51.0 58.5 91.6 55.5** |       |

VSRN\textsuperscript{[10]} is one of the machine annotators used to construct the ECCV Caption dataset.

\textsuperscript{[10]} Li et al., Visual Semantic Reasoning for Image-text Matching, ICCV 2019.
**Experiments: Ablation Study on Flickr30K**

| Similarity       | Arch.  | RSUM  |
|------------------|--------|-------|
| MIL[11]          | Ours   | 491.7 |
| MP[12]           | Ours   | 490.5 |
| Ours (Chamfer)   | Ours   | 499.6 |
| Ours (S-Chamfer) | PIE-Net| 483.3 |
| Ours (S-Chamfer) | Ours   | **500.8** |

**Impact of set-similarity metric**

*Smooth-Chamfer similarity is best suited to our framework.*

| Setting               | log(Var.) | RSUM  |
|-----------------------|-----------|-------|
| PIE-Net[11,12]        | -7.35     | 483.3 |
| Ours \w MP            | -5.27     | 490.5 |
| Transformer[13]       | -2.27     | 496.1 |
| Ours                  | -2.13     | **500.8** |

**Impact of set-embedding architecture**

*Our architecture results in most diverse embeddings and best performance.*

Circular variance $\text{Var} = 1 - \left\| \sum_{e \in S} \frac{e}{|S|} \right\|_2$

[11] Song and Soleymani, Polysemous Visual Semantic Embedding for Cross-modal Retrieval, CVPR 2019.
[12] Chun et al., Probabilistic Embeddings for Cross-modal Retrieval, CVPR 2021.
[13] Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021.
Experiments: Ablation Study on Flickr30K

| $S^Y(1)$ | $S^Y(2)$ | $S^Y(3)$ | $S^Y(4)$ | RSUM |
|----------|----------|----------|----------|------|
| ✓        | ✓        | ✓        | ✓        | 500.8|
| ✓        |          |          |          | 491.1|
| ✓        |          |          |          | 309.6|
| ✓        |          |          |          | 484.9|
|          |          |          | ✓        | 486.0|

| $S^T(1)$ | $S^T(2)$ | $S^T(3)$ | $S^T(4)$ | RSUM |
|----------|----------|----------|----------|------|
| ✓        | ✓        | ✓        | ✓        | 500.8|
| ✓        |          |          |          | 481.9|
|          |          |          |          | 483.0|
|          |          |          |          | 481.7|
|          |          |          | ✓        | 497.2|

Contribution of each embedding element
Experiments: Qualitative Examples

R1: Picture of an outdoor place that is very beautiful.

R1: An old countryle store has a display of stuffed animals outside.

R1: A park is full of patrons on a fall day.

R1: A country store with several teddy bears and geese there.

R1: Here is a soul in the image alone.

R1: A man in a robe eating a chocolate donut.

R1: A hairy man eating a chocolate doughnut in his house.

R1: A man is holding a chocolate dessert in his hand as he stares ahead.
Conclusion

• Contributions
  • A new set-based embedding architecture
  • A new set similarity metric
  • Outstanding performance on four public benchmarks

• Next on agenda
  • Adopting CLIP-pretrained weights\textsuperscript{[14]}
  • Adopting an advanced slot attention mechanism (\textit{e.g.}, \textsuperscript{[15]})
  • Learning vision-language models with the proposed method

\textsuperscript{[14]} Radford \textit{et al.}, Learning Transferable Visual Models From Natural Language Supervision, ICML 2021.
\textsuperscript{[15]} Kim \textit{et al.}, Shatter and Gather: Learning Referring Image Segmentation with Text Supervision, ICCV 2023.
References

[1] Locatello et al., Object-centric Learning with Slot Attention, NeurIPS 2020.
[2] Gretton et al., A Kernel Two-sample Test, JMLR 2012.
[3] Lin et al., Microsoft COCO: Common Objects in Context, ECCV 2014.
[4] Plummer et al., Flickr30k Entities: Collecting Region-to-phrase Correspondences for Richer Image-to-sentence Models, ICCV 2015.
[5] Chun et al., ECCV Caption, Correcting False Negatives by Collecting Machine-and-human-verified Image-Caption Associations for MS-COCO, ECCV 2022.
[6] Parekh et al., Crisscrossed Captions, EACL 2020.
[7] Jiacheng et al., Learning the Best Pooling Strategy for Visual Semantic Embedding, CVPR 2021.
[8] Zhang et al., Negative-aware Attention Framework for Image-text Matching., CVPR 2022.
[9] Lee et al., Stacked Cross Attention for Image-text Matching, ECCV 2018.
[10] Li et al., Visual Semantic Reasoning for Image-text Matching, ICCV 2019.
[11] Song and Soleymani, Polysemous Visual Semantic Embedding for Cross-modal Retrieval, CVPR 2019.
[12] Chun et al., Probabilistic Embeddings for Cross-modal Retrieval, CVPR 2021.
[13] Dosovitskiy et al., An Image is Worth 16x16 Words, ICLR 2021.
[14] Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021.
[15] Kim et al., Shatter and Gather: Learning Referring Image Segmentation with Text Supervision, ICCV 2023.
