Ladder Loss for Coherent Visual-Semantic Embedding

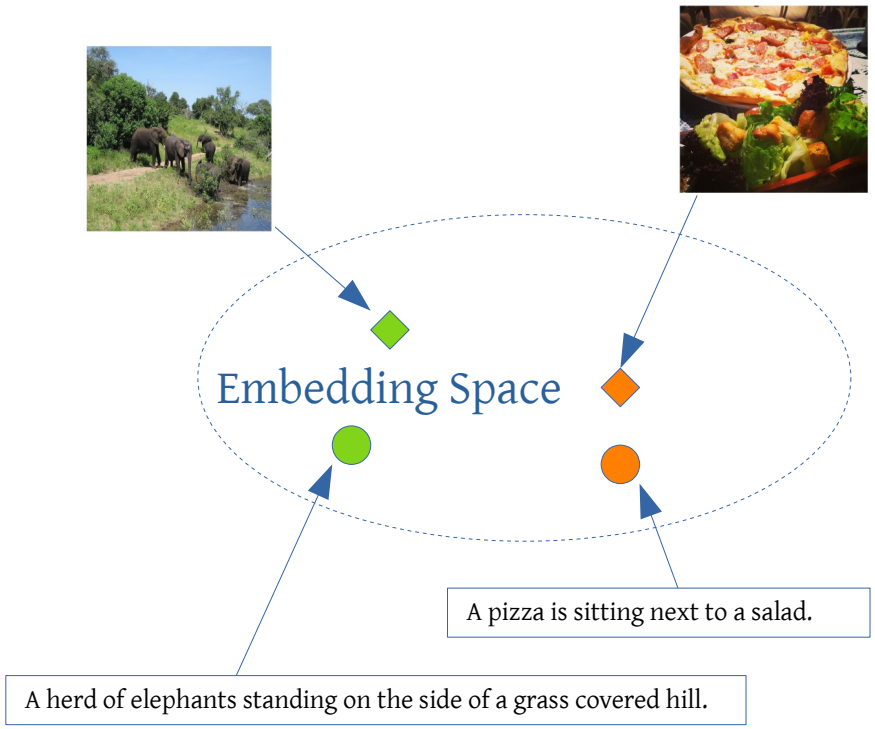
(AAAI-2020, Feb 7-12, New York, NY, USA)

Mo Zhou
Xidian University
Dec 25 2019

https://arxiv.org/abs/1911.07528

Visual-Semantic Embedding

- Cross-Modal Retrieval (image ←→ text)



WHAT

* multi-modal joint embedding
* Map images & texts into a common space
* Images → Convolutional Neural Nets
* Text → Recurrent Neural Nets

WHY

* Retrieval by comparing embedding distance
* Image/text generation using embedding vecs
* Visual question answering

Visual-Semantic Embedding

- How

  (1) Canonical Correlation Analysis (CCA)

  (2) Deep metric learning-based embedding space learning

- Deep metric learning-based methods scale better to large scale datasets and achieve better performance

Deep Metric Learning

- Metric

Given a set X, a metric is a (distance function) $d : X \times X \mapsto [0, \infty)$

Visual-Semantic Embedding

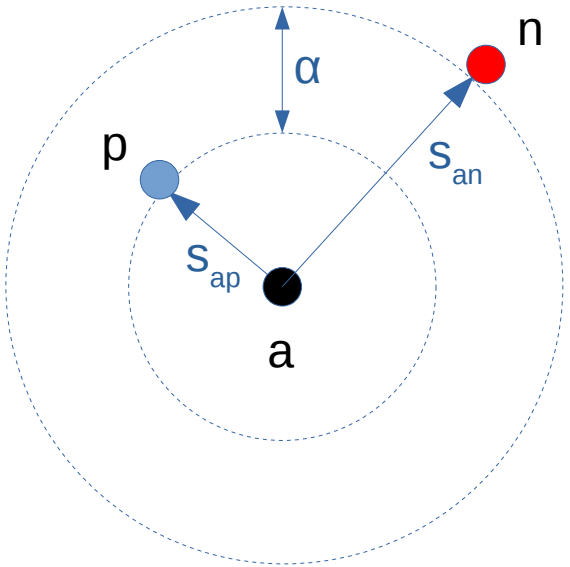
Image & Text embedding vectors $X = \{x_{\text{image},1}, x_{\text{text},1}, \ldots, x_{\text{image},N}, x_{\text{text},N}\}$

Subtle problem: Loss functiom for DML

* Contrastive loss (restrictive)
* Triplet loss (less-restrictive)
* Variants of Triplet loss, e.g. PDDM, Histogram Loss, quadruplets.
* n-pair loss
* lift-structure loss

Existing Problem

- Relevance → Binary Variable. e.g. Triplet loss

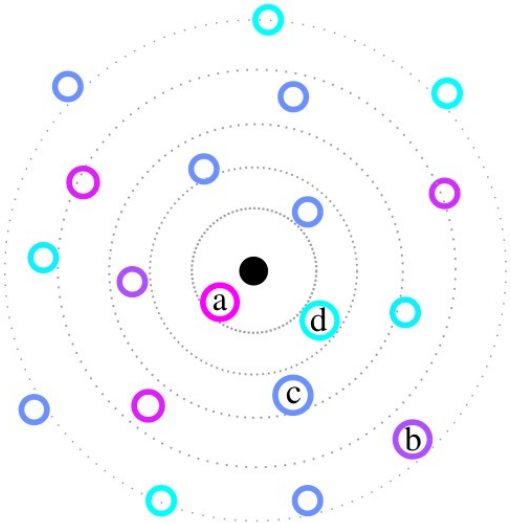


$$s(v_q, h_q) - s(v_q, h_p) > \alpha$$

$$L_{tri}(q) = \sum_{p \in \mathcal{N}^{-q}} [\alpha - s(v_q, h_q) + s(v_q, h_p)]_+,$$

Existing Problem

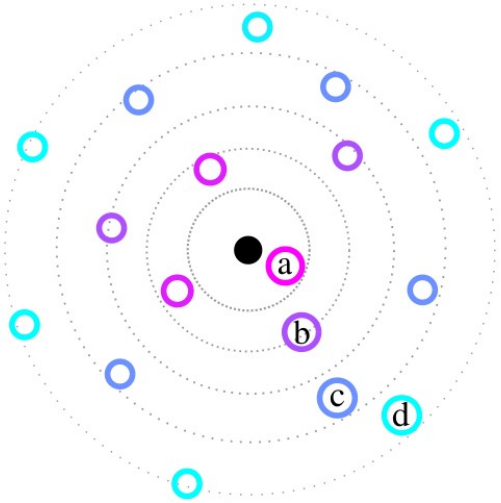
(a) A herd of elephants standing on the side of a grass covered hill.
(b) An elephant stands near water and a stone wall.
(c) A pair of elephants with their trunks entwined.
(d) A group of four skiers posing for a picture.



Retrieval Result: (a)→(d)→(c)→(b)

Typical

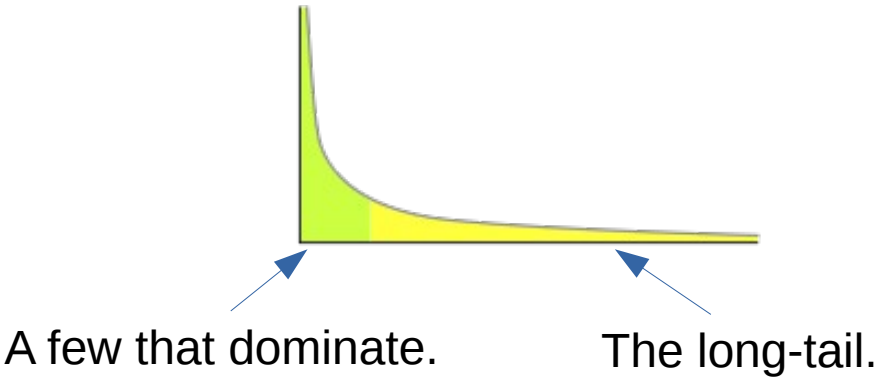
Ideal



Retrieval Result: (a)→(b)→(c)→(d)

Long-Tail Query

- Where a matching candidate may not necessarily exist in the database.



Typical embedding models will fail to retrieve the “most relevant” candidates.

Our Approach

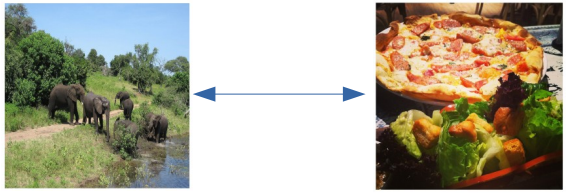
- Relevance Degree (Continuous Variable)
- Ladder Loss (Extended form of Triplet Loss)
- Coherent Score (Evaluate Ranking Coherence)

Relevance Degree

Relevance?

A pizza is sitting next to a salad.

Ideal ground-truth: Human Annotation | Combinatorial Explosion



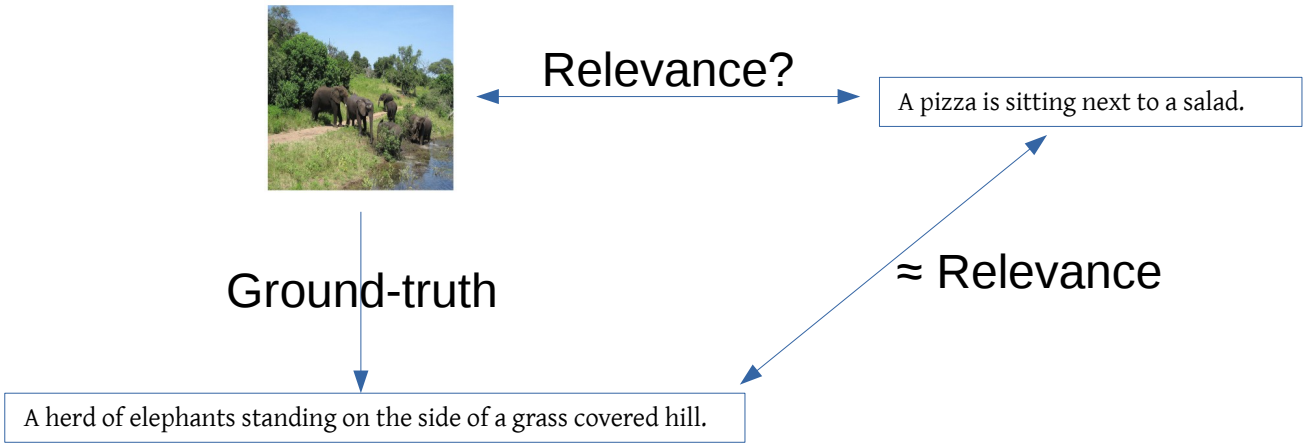
Single-modal (image) similarity.

A pizza is sitting next to a salad.

A herd of elephants standing on the side of a grass covered hill.

Single-modal (text) similarity.

Relevance Degree

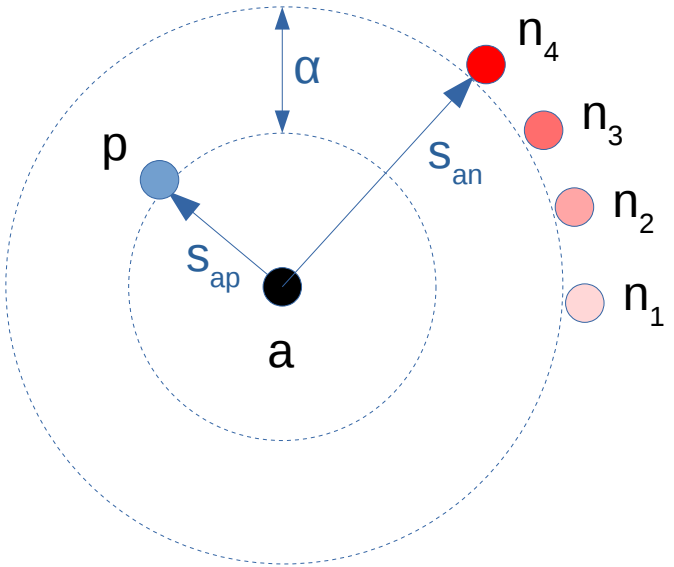


$$R(i_q, t_p) = R(t_q, t_p) = \mathrm{BERT}(t_q, t_p).$$

BERT achieves 0.88 in pearson correlation on STS-B.

Ladder Loss

- Triplet Loss



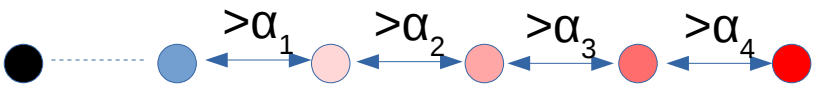
$$s(v_q, h_q) - s(v_q, h_p) > \alpha$$

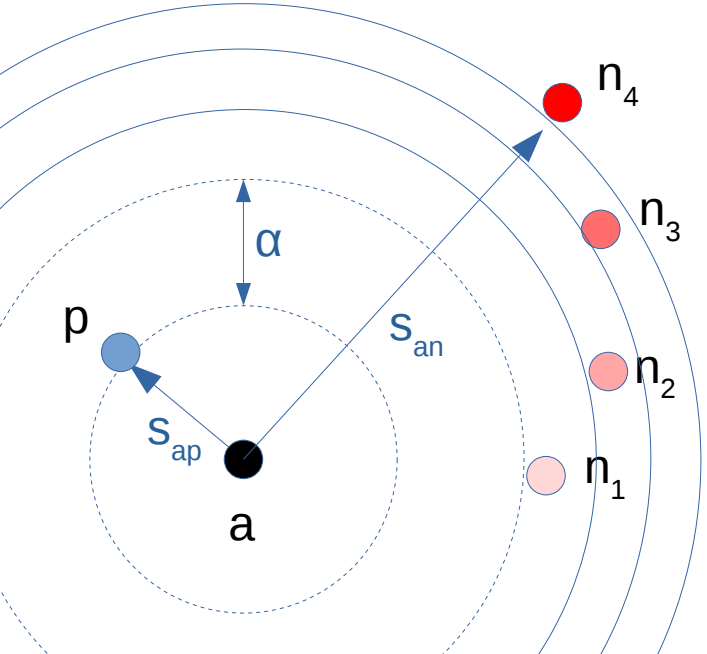
Problem:
treats candidates in a bipolar way.

Goal:
leverage the continuous relevance degree and learn coherent embedding space.

Ladder Loss

- Ladder Loss



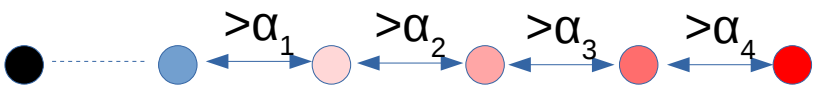


$$s(i_q, t_q) > s(i_q, t_{n1}) > s(i_q, t_{n2}) > \ldots > s(i_q, t_{nl})$$

$$\begin{aligned}
&s(v_q, h_q) - s(v_q, h_i) > \alpha_1, (i \in \mathcal{N}_1^{-q}),\\
&s(v_q, h_i) - s(v_q, h_j) > \alpha_2, (i \in \mathcal{N}_1^{-q}, j \in \mathcal{N}_2^{-q}),\\
&s(v_q, h_j) - s(v_q, h_k) > \alpha_3, (j \in \mathcal{N}_2^{-q}, k \in \mathcal{N}_3^{-q}),\\
&\ldots,
\end{aligned}$$

Ladder Loss

- Ladder Loss



$$L_{lad}(q) = \beta_1 L^1_{lad}(q) + \beta_2 L^2_{lad}(q) + \beta_3 L^3_{lad}(q), \quad (6)$$

$$L^1_{lad}(q) = \sum\nolimits_{i \in \mathcal{N}^{-q}_{1:L}} [\alpha_1 - s(v_q, h_q) + s(v_q, h_i)]_+,$$

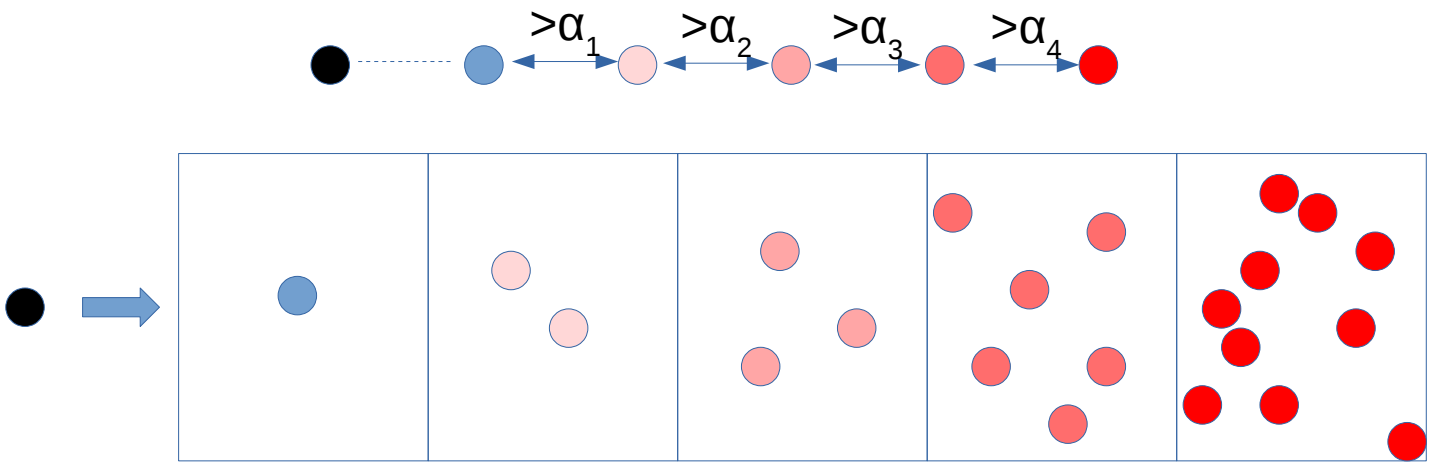
$$L^2_{lad}(q) = \sum\nolimits_{i \in \mathcal{N}^{-q}_{1}, j \in \mathcal{N}^{-q}_{2:L}} [\alpha_2 - s(v_q, h_i) + s(v_q, h_j)]_+, \quad (7)$$

$$L^3_{lad}(q) = \sum\nolimits_{j \in \mathcal{N}^{-q}_{2}, k \in \mathcal{N}^{-q}_{3:L}} [\alpha_3 - s(v_q, h_j) + s(v_q, h_k)]_+,$$

β： Balancing factor.

Ladder Loss

- Hard Contrastive Sampling

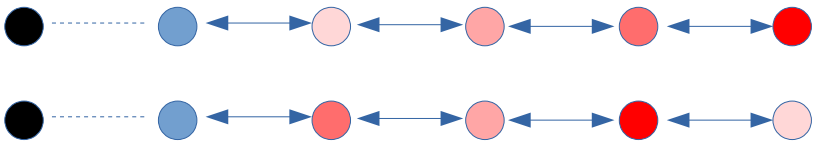


Number of inequality chain: combinatorial explosion.

Coherent Score

Ground-truth:

Model output:



(How Coherent?)

* Traditional performance metric Recall @ K: Only cares about the rank of

* Coherent Score: Kendall's Ranking Correlation Coefficient τ

Experiments

- MS-COCO (120k images, each with 5 annotations)
- Flickr30k (30k images, each with 5 annotations)

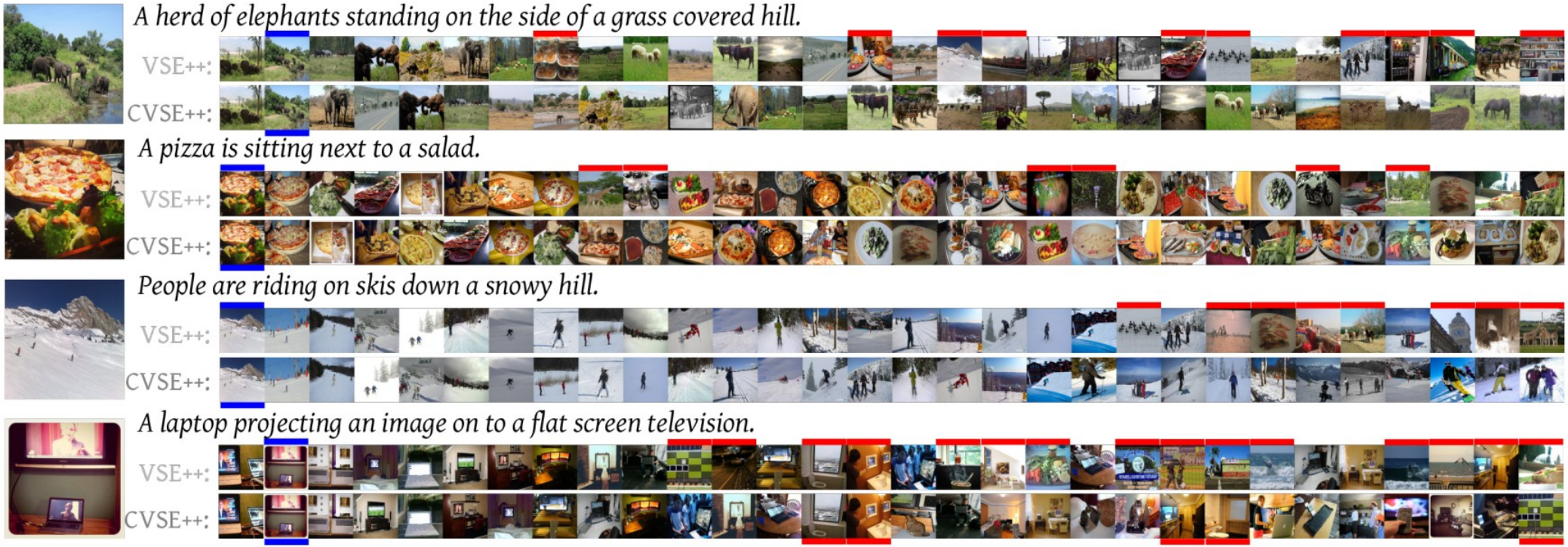
- BERT STS Acceleration

  CBoW < threshold → relevance degree (coarse)

  CBoW > threshold → relevance degree (fine)

  Hybrid model with threshold set to 0.8 achieves 0.79 in pearson correlation on STS-B.

Visualization (T→ I)



Visualization (I→ T)

VSE++:

A herd of elephants on the grassy plains.
A herd of elephants standing on the side of a grass covered hill.
Large group of elephants with men dressed in blue riding them.
An elephant drinking water while the rest of the herd is walking in ...
A group of men riding on top of elephants.
A large elephant is shown walking through the terrain.
A group of four skiers posing for a picture.
An elephant standing on top of a lush green hillside.
A cluster of small boats in shallow water.
A herd of cows is in the meadow.
a pair of elephants with their trunks entwined
Zebras and wort hogs living together on the plains.
three black cows are standing in a field
A group of stuffed bears are arranged in a display.
a variety of old school show cars lined in a row
a long buffet length table covered in many dishes of different foods
cherry tomatoes and various food dishes on a table top
there are many cars that are waiting for the light to turn green
Three plates of toasted sandwiches are on a counter.
A compact car is parked beside many motorcycles.
Three pizzas are sitting in the windows a small red car.
There are two toilets and a sink in this bathroom.
A gray couch sitting on top of a rug in a living room.
A living room is shown with tables and seating.
An elephant stands near water and a stone wall.
A busy intersection filled with traffic under traffic lights.
Three zebras stand camouflaged in the wide open plains.
Four giraffes are standing in front of a small puddle.
Many cars in a desert area with planes in the back.
A living room with ample lighting from windows and lamps and brown ...

CVSE++:

A herd of elephants standing on the side of a grass covered hill.
A herd of elephants on the grassy plains.
An elephant drinking water while the rest of the herd is walking in ...
Large group of elephants with men dressed in blue riding them.
A large elephant is shown walking through the terrain.
An elephant standing on top of a lush green hillside.
A group of men riding on top of elephants.
a lawn cut out of a big elephant in a yard outside
A herd of cattle is shown in an open field next to a river.
An elephant stands near water and a stone wall.
An elephant with a red cloth over it.
a pair of elephants with their trunks entwined
Four giraffes are standing in front of a small puddle.
Zebras and wort hogs living together on the plains.
a number of cows in a field near a tree
A man on a horse corralling a herd of cattle.
three black cows are standing in a field
A group of birds drinking from a puddle
Many sheep and goats cross a road between grassy fields.
An elephant walks behind another elephant holding it's tail with it's ...
A couple of animals that are in the grass.
there are two very tall giraffes that are in the wild
A herd of cows is in the meadow.
THREE ZEBRA ARE STANDING TOGETHER IN THE OPEN LAND
Men on horses are driving many other horses in a group.
A group of stuffed bears are arranged in a display.
Horses walking in meadow with trees and buildings in background.
Two red and white cows in a pasture with a pond.
Several cows standing in a row on a large open farm.
A couple of bears sitting on a coach, one is holding a book.

VSE++:

a pizza is sitting next to a salad
Three pizzas are sitting in the windows a small red car.
Two photos, each with pizzas, one ham and cheese, one supreme.
A black pan holds several small, cooked pizzas.
A pan filled with meat and vegetables cooking on a stove top.
A slice of pizza with a bunch of vegetable toppings
A pizza and tea kettle sitting on a stove.
the pizza has cheese, tomato sauce, and olives
A pizza with two slices missing from it.
a person is eating a pizza with a fork
A homemade pizza is waiting to be cooked.
Two pieces of bread with sauce on them next to a bowl of chicken ...
A thin crust pizza is topped with olives and onions.
This meal has beans and three slices of pita bread.
a fresh baked pizza with piping hot cheese
A woman puts spices on a skillet of food that is cooking.
an image of two people on skiis going down the slope
A person cutting a pizza on top of parchment paper.
A white plate topped with a vegetables salad.
A herd of elephants standing on the side of a grass covered hill.
An arrangement or orchids and greenery in a flower pot.
A loaded open sandwich at a dinner table.
A plate with a hamburger and a salad on it
a hot dog topped with tomatoes and peppers with a side of fries
there are many fire trucks and=f fire cars in this garage
A child's bedroom has plenty of storage space and a small chair.
A toilet is in a tiled bathroom with a large window looking out on ...
Man about to bite into hot dog with many toppings.
Several buses and cars sitting at a traffic light on a road.
A kitchen with a stove, microwave, telephone, table and chair.

CVSE++:

a pizza is sitting next to a salad
A black pan holds several small, cooked pizzas.
Two photos, each with pizzas, one ham and cheese, one supreme.
A homemade pizza is waiting to be cooked.
A pan filled with meat and vegetables cooking on a stove top.
a fresh baked pizza with piping hot cheese
the pizza has cheese, tomato sauce, and olives
Two pieces of bread with sauce on them next to a bowl of chicken ...
A slice of pizza with a bunch of vegetable toppings
A thin crust pizza is topped with olives and onions.
A pizza and tea kettle sitting on a stove.
Three pizzas are sitting in the windows a small red car.
A pizza with two slices missing from it.
a person is eating a pizza with a fork
A very tasty looking dish with some broccoli and other items.
A plate with a hamburger and a salad on it
A person cutting a pizza on top of parchment paper.
a meat dish with broccoli and peppers on a white plate
A white plate topped with a vegetables salad.
Two meat sandwiches with potato salad on a white plae
A woman puts spices on a skillet of food that is cooking.
A loaded open sandwich at a dinner table.
A plate of hotdogs, lettuce and tomatoes next to a cup of beer.
a table with many different vegetables and fruit
A metal platter topped with fish and vegetables.
A simple meal of chicken and a large serving of broccoli.
Two small trays filled with an assortment of food items
a cheese pizza cut into slices on a table
This meal has beans and three slices of pita bread.
cherry tomatoes and various food dishes on a table top

Experiments: MS-COCO

MS-COCO (1000 Test Samples)												
Model	Image→Sentence						Sentence→Image					
	CS@100	CS@1000	Mean R	R@1	R@5	R@10	CS@100	CS@1000	Mean R	R@1	R@5	R@10
Random	0.018	0.009	929.9	0.0	0.3	0.5	0.044	0.005	501.0	0.1	0.5	0.9
VSE++ (VGG19)	0.235	0.057	5.7	56.7	83.9	92.0	0.237	0.057	9.1	42.6	76.5	86.8
CVSE++ (VGG19)	0.256	0.347	4.1	56.8	83.6	92.2	0.257	0.223	7.3	43.2	77.5	88.1
VSE++ (VGG19,FT)	0.253	0.047	2.9	62.5	88.2	95.2	0.246	0.042	6.5	49.9	82.8	91.2
CVSE++ (VGG19,FT)	0.256	0.419	2.8	63.2	89.9	95.0	0.251	0.287	5.3	50.5	83.6	92.8
VSE++ (Res152)	0.238	0.079	2.8	63.2	88.9	95.5	0.236	0.080	7.3	47.4	80.3	89.9
CVSE++ (Res152)	0.265	0.358	2.8	66.7	90.2	94.0	0.256	0.236	6.1	48.4	81.0	90.0
VSE++ (Res152,FT)	0.241	0.071	2.4	68.0	91.9	97.4	0.239	0.068	6.3	53.5	85.1	92.5
CVSE++ (Res152,FT)	0.265	0.446	2.4	69.1	92.2	96.1	0.255	0.275	4.7	55.6	86.7	93.8

MS-COCO (5000 Test Samples)												
Model	Image→Sentence						Sentence→Image					
	CS@500	CS@5000	Mean R	R@1	R@5	R@10	CS@500	CS@5000	Mean R	R@1	R@5	R@10
VSE++ (Res152)	0.227	0.078	10.6	36.3	66.8	78.7	0.224	0.084	30.9	25.6	54.0	66.9
CVSE++ (Res152)	0.253	0.354	9.7	39.3	69.1	80.3	0.246	0.239	25.2	25.8	54.0	67.3
VSE++ (Res152,FT)	0.231	0.073	7.7	40.2	72.5	83.3	0.228	0.073	25.1	30.7	60.7	73.3
CVSE++ (Res152,FT)	0.255	0.439	7.4	43.2	73.5	84.1	0.242	0.280	18.6	32.4	62.2	74.6

Table 1: Comparison between VSE++ and CVSE++ in terms of CS@K and R@K on MS-COCO.

Experiments: Flickr30k

Model	Image→Sentence						Sentence→Image					
	CS@100	CS@1000	Mean R	R@1	R@5	R@10	CS@100	CS@1000	Mean R	R@1	R@5	R@10
Random	0.02	-0.005	988.3	0.0	0.3	0.4	-0.033	-0.003	503.0	0.2	0.6	1.1
VSE++ (VGG19)	0.116	0.139	18.2	40.7	68.4	78.0	0.115	0.124	26.9	28.7	58.6	69.8
CVSE++ (VGG19)	0.129	0.255	16.4	42.8	69.2	78.9	0.127	0.144	26.4	29.0	59.2	71.1
VSE++ (VGG19,FT)	0.128	0.130	14.7	44.6	73.3	82.0	0.125	0.110	22.8	31.9	63.0	74.5
CVSE++ (VGG19,FT)	0.133	0.260	13.0	44.8	73.1	82.3	0.131	0.160	20.8	33.8	63.9	75.1
VSE++ (Res152)	0.126	0.127	10.2	49.3	78.9	86.4	0.115	0.112	20.0	35.9	65.9	75.6
CVSE++ (Res152)	0.133	0.247	9.3	50.2	78.8	87.3	0.120	0.147	20.0	37.1	66.9	76.4
VSE++ (Res152,FT)	0.130	0.122	7.8	54.1	81.0	88.7	0.122	0.114	16.2	39.8	70.0	79.0
CVSE++ (Res152,FT)	0.141	0.273	7.4	56.6	82.5	90.2	0.126	0.172	15.7	42.4	71.6	80.8

Table 2: Comparison between VSE++ and CVSE++ in terms of CS@K and R@K on Flickr30K.

Discussion: Balancing Factor

β_2	Image→Sentence						Sentence→Image					
	CS@100	CS@1000	Mean R	R@1	R@5	R@10	CS@100	CS@1000	Mean R	R@1	R@5	R@10
0.0	0.238	0.079	2.8	63.2	88.9	95.5	0.236	0.08	7.3	47.4	80.3	89.9
0.25	0.265	0.358	2.8	66.7	90.2	94.0	0.256	0.236	6.1	48.4	81.0	90.0
1.0	0.266	0.417	3.9	64.0	88.2	93.1	0.259	0.264	6.2	47.4	79.0	88.9

Table 3: Performance of the proposed CVSE++(Res152) with respect to the parameter β_2 (On MS-COCO dataset).

Discussion: Scope-of-Interest

L	Image→Sentence							Sentence→Image						
	CS@100	CS@200	CS@1000	Mean R	R@1	R@5	R@10	CS@100	CS@200	CS@1000	Mean R	R@1	R@5	R@10
1	0.238	0.188	0.079	2.8	63.2	88.9	95.5	0.236	0.189	0.08	7.3	47.4	80.3	89.9
2	0.265	0.252	0.358	2.8	66.7	90.2	94.0	0.256	0.253	0.236	6.1	48.4	81.0	90.0
3	0.267	0.274	0.405	3.2	65.7	89.3	94.1	0.261	0.258	0.244	6.3	48.4	80.3	89.4

Table 4: Performance of the proposed CVSE++(Res152) with respect to the ladder number L. (On MS-COCO dataset)

Conclusion & Future Work

- Continuous Relevance Degree
- New Ladder Loss Function for Coherent Embedding
- Coherent Score as performance metric
- Our method → Coherent Visual-Semantic Emb Space

- Future work: Extend to other metric learning apps.

Thanks!

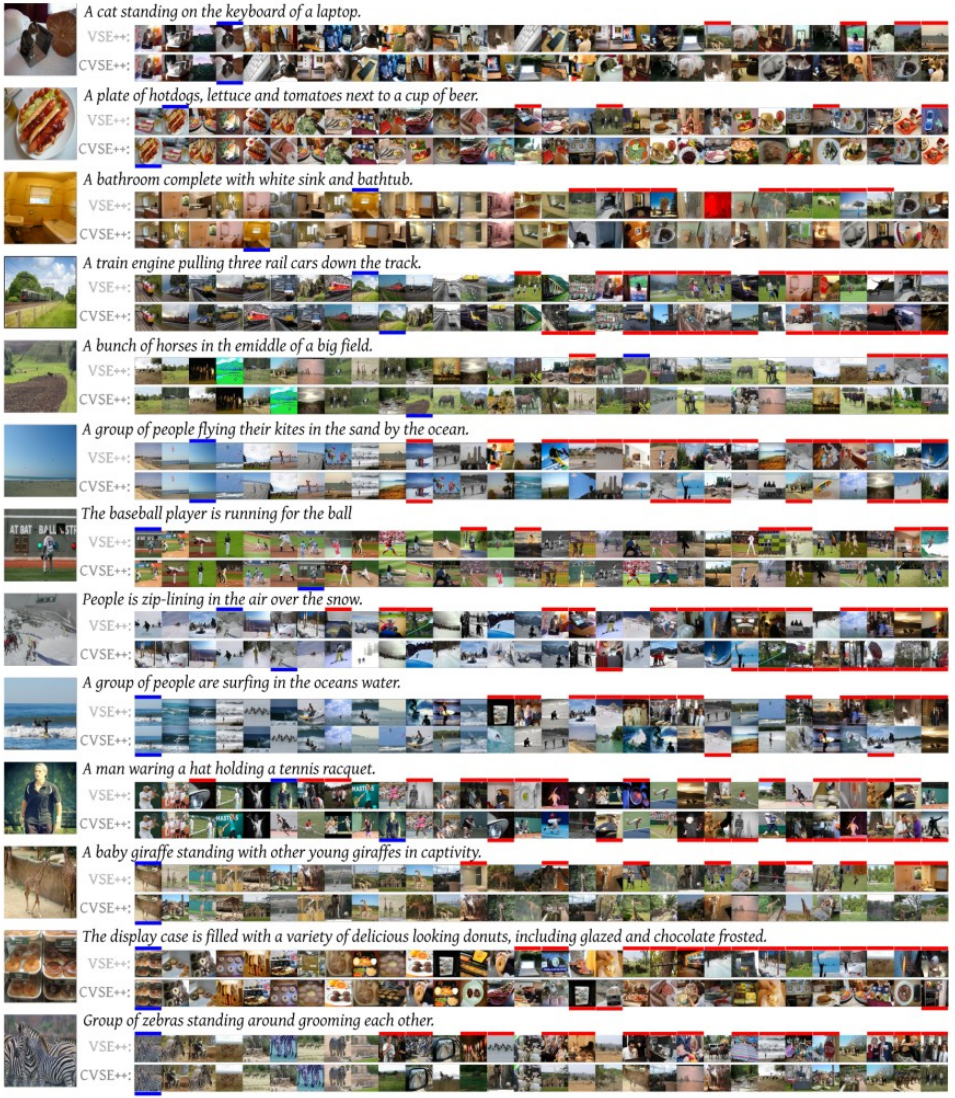
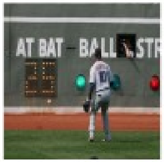


Figure 2. Comparison of the sentence-to-image top-30 retrieval results between VSE++ (baseline, 1st row) and CVSE++ (Ours, 2nd row). For each query sentence, the ground-truth image is shown on the left, the totally-relevant and totally-irrelevant retrieval results are marked by blue and red overlines/underlines, respectively. Figure best viewed in color.

VSE++:

CVSE++:



VSE++:

CVSE++:

VSE++:

CVSE++:

VSE++:

CVSE++:

VSE++:

CVSE++:

VSE++:

CVSE++:

VSE++:

CVSE++:

VSE++:

CVSE++:

VSE++:

CVSE++: