Ladder Loss for Coherent Visual-Semantic Embedding

Mo Zhou,[1] **Zhenxing Niu,**[2] **Le Wang,**[3]* **Zhanning Gao,**[2,3] **Qilin Zhang,**[4] **Gang Hua**[5]

[1]Xidian University, [2]Alibaba Group, [3]Xi'an Jiaotong University, [4]HERE Technologies, [5]Wormpex AI Research

{cdluminate, zhenxingniu, zhanninggao, samqzhang, ganghua}@gmail.com, lewang@xjtu.edu.cn

Abstract

For visual-semantic embedding, the existing methods normally treat the relevance between queries and candidates in a bipolar way – relevant or irrelevant, and all "irrelevant" candidates are uniformly pushed away from the query by an equal margin in the embedding space, regardless of their various proximity to the query. This practice disregards relatively discriminative information and could lead to suboptimal ranking in the retrieval results and poorer user experience, especially in the long-tail query scenario where a matching candidate may not necessarily exist. In this paper, we introduce a continuous variable to model the relevance degree between queries and multiple candidates, and propose to learn a coherent embedding space, where candidates with higher relevance degrees are mapped closer to the query than those with lower relevance degrees. In particular, the new ladder loss is proposed by extending the triplet loss inequality to a more general inequality chain, which implements variable push-away margins according to respective relevance degrees. In addition, a proper Coherent Score metric is proposed to better measure the ranking results including those "irrelevant" candidates. Extensive experiments on multiple datasets validate the efficacy of our proposed method, which achieves significant improvement over existing state-of-the-art methods.

Introduction

Visual-semantic embedding aims to map images and their descriptive sentences into a common space, so that we can retrieve sentences given query images or vice versa, which is namely cross-modal retrieval (Ji et al. 2017). Recently, the advances in deep learning have made significant progress on visual-semantic embedding (Kiros, Salakhutdinov, and Zemel 2014; Karpathy and Fei-Fei 2015; Karpathy, Joulin, and Fei-Fei 2014; Faghri et al. 2018). Generally, images are represented by the Convolutional Neural Networks (CNN), and sentences are represented by the Recurrent Neural Networks (RNN). A triplet ranking loss is subsequently optimized to make the corresponding representations as close as possible in the embedding space (Schroff, Kalenichenko, and Philbin 2015; Sohn 2016).
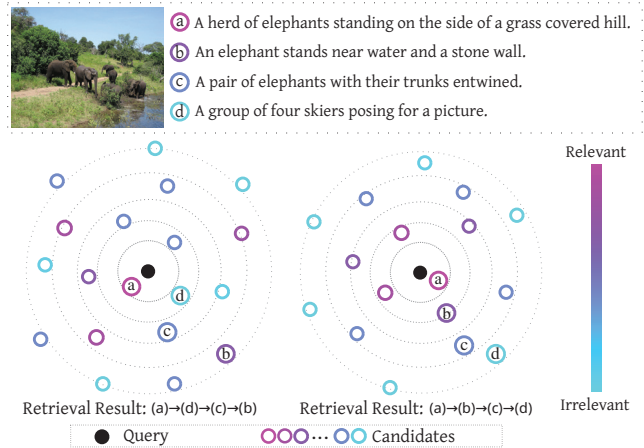
*Corresponding author.

Figure 1: Comparison between the incoherent (left) and coherent (right) visual-semantic embedding space. Existing methods (left) pull the totally-relevant sentence (a) close to the query image, while pushing away all other sentences (b, c, and d) equally. Therefore, the relative proximity of (b, c, and d) are not necessarily consistent with their relevance degrees to the query (solid black dot). On contrary, our approach (right) explicitly preserves the proper relevance order in the retrieval results.

For visual-semantic embedding, previous methods (Hadsell, Chopra, and LeCun 2006; Schroff, Kalenichenko, and Philbin 2015) tend to treat the relevance between queries and candidates in a bipolar way: for a query image, only the corresponding ground-truth sentence is regarded as **relevant**, and other sentences are *equally* regarded as **irrelevant**. Therefore, with the triplet ranking loss, only the relevant sentence is pulled close to the query image, while all the irrelevant sentences are pushed away *equally*, *i.e.*, be pushed from the query by an equal margin. However, among those so-called **irrelevant** sentences, some are more relevant to the query than others, thus should be treated accordingly.

Similarly, it is arguably a disadvantage in recent retrieval evaluation metrics which disregard the ordering/ranking of retrieved "irrelevant" results. For example, the most popular Recall@K (*i.e.*, R@K) (Kiros, Salakhutdinov, and Zemel 2014; Karpathy and Fei-Fei 2015; Faghri et al. 2018) is

purely based on the ranking position of the ground-truth candidates (denoted as *totally-relevant* candidates in this paper); while *neglecting* the ranking order of all other candidates. However, the user experience of a practical cross-modal retrieval system could be heavily impacted by the ranking order of all top-N candidates, including the "irrelevant" ones, as it is often challenging to retrieve enough totally-relevant candidates in the top-N results (known as the long-tail query challenge (Downey, Dumais, and Horvitz 2007)). Given a query from the user, when a exact matching candidate does not exist in the database, a model trained with only bipolar supervision information will likely fail to retrieve those somewhat relevant candidates, and produce a badly ordered ranking result. As demonstrated in Fig. 1, given a query image (solid black dot), the ground-truth sentence (a) is the totally-relevant one, which does occupy the top of the retrieved list. Besides that, the sentence (b) is notably more relevant than (c) or (d), so ideally the (b) should be ranked before the (c), and the (d) should be ranked at the bottom.

Therefore, it is beneficial to formulate the semantic **relevance degree** as a continuous variable rather than a binary variable (*i.e.*, relevant or irrelevant). And the relevance degree should be incorporated into embedding space learning, so that the candidates with higher relevance degrees will be closer to the query than those with lower degrees.

In this paper, we first propose to measure the relevance degree between images and sentences, based on which we design the **ladder loss** to learn a *coherent* embedding space. The "coherent" means that the similarities between queries and candidates are conformal with their relevance degrees. Specifically, the similarity between the query image i_q and its totally-relevant sentence t_q in the conventional triplet loss (Faghri et al. 2018) is encouraged to be greater than the similarity between the i_q and other sentences t_p. Likewise, with the ladder loss formulation, we consider the relevance degrees of all sentences, and extend the inequality $s(i_q, t_q) > s(i_q, t_p)$ to an inequality chain, *i.e.*, $s(i_q, t_q) > s(i_q, t_{p_1}) > s(i_q, t_{p_2}) > \cdots > s(i_q, t_{p_L})$, where t_{p_l} is more relevant to i_q than $t_{p_{l+1}}$, and $s(\cdot, \cdot)$ denotes cosine similarity. Using the inequality chain, we design the ladder loss so that the sentences with lower relevance degrees will be pushed away by a larger margin than the ones with higher relevance degrees. As a result, it leads to learn a coherent embedding space, and both the totally-relevant as well as the somewhat-relevant sentences can be properly ranked.

In order to better evaluate the quality of retrieval results, we propose a new **Coherent Score (CS)** metric, which is designed to measure the alignment between the real ranking order and the expected ranking order. The expected ranking order is decided according to the relevance degrees, so that the CS can properly reflect user experience for cross-modal retrieval results. In brief, our contributions are:

1. We propose to formulate the relevance degree as a continuous rather than a binary variable, which leads to learn a coherent embedding space, where both the totally-relevant and the somewhat-relevant candidates can be retrieved and ranked in a proper order.

2. To learn a coherent embedding space, a ladder loss is pro-

posed by extending the inequality in the triplet loss to an inequality chain, so that candidates with different degrees will be treated differently.

3. A new metric, Coherent Score (CS), is proposed to evaluate the ranking results, which can better reflect user experience in a cross-modal retrieval system.

Related Work

Visual-semantic Embedding, as a kind of multi-modal joint embedding, enables a wide range of tasks in image and language understanding, such as image-caption retrieval (Karpathy, Joulin, and Fei-Fei 2014; Kiros, Salakhutdinov, and Zemel 2014; Faghri et al. 2018), image captioning, and visual question-answering (Malinowski, Rohrbach, and Fritz 2015). Generally, the methods of visual-semantic embedding could be divided into two categories. The first category is based on Canonical Correlation Analysis (CCA) (Hardoon, Szedmak, and Shawe-Taylor 2004; Gong et al. 2014a; 2014b; Klein et al. 2014) which finds linear projections that maximize the correlation between projected vectors from the two modalities. Extensions of CCA to a deep learning framework have also been proposed (Andrew et al. 2013; Yan and Mikolajczyk 2015).

The second category involves metric learning-based embedding space learning (Frome et al. 2013; Wang, Li, and Lazebnik 2016; Faghri et al. 2018). DeViSE (Frome et al. 2013; Socher et al. 2014) learns linear transformations of visual and textual features to the common space. After that, Deep Structure-Preserving (DeepSP) (Wang, Li, and Lazebnik 2016) is proposed for image-text embedding, which combines cross-view ranking constraints with within-view neighborhood structure preservation. In (Niu et al. 2017), Niu *et al.* propose to learn a hierarchical multimodal embedding space where not only full sentences and images but also phrases and image regions are mapped into the space. Recently, Fartash *et al.* (Faghri et al. 2018) incorporate hard negatives in the ranking loss function, which yields significant gains in retrieval performance. Compared to CCA-based methods, metric learning-based methods scale better to large dataset with stochastic optimization in training.

Metric learning, has many other applications such as face recognition (Schroff, Kalenichenko, and Philbin 2015) and fine-grained recognition (Oh Song et al. 2016; Wu et al. 2017; Yuan, Yang, and Zhang 2017). The loss function design in metric learning could be a subtle problem. For example, the contrastive loss (Hadsell, Chopra, and LeCun 2006) pulls all positives close, while all negatives are separated by a fixed distance. However, it could be severely restrictive to enforce such fixed distance for all negatives. This motivated the triplet loss (Schroff, Kalenichenko, and Philbin 2015), which only requires negatives to be farther away than any positives on a per-example basis, *i.e.*, a less restrictive relative distance constraint. After that, many variants of triplet loss are proposed. For example, PDDM (Huang, Loy, and Tang 2016) and Histogram Loss (Ustinova and Lempitsky 2016) use quadruplets. Beyond that, the n-pair loss (Sohn 2016) and Lifted Structure (Oh Song et al. 2016) define constraints on all images in a batch. However, all the aforemen-

tioned methods formulate the relevance as a binary variable. Thus, our ladder loss could be used to boost those methods.

Our Approach

Given a set of image-sentence pairs $\mathcal{D} = \{(i_n, t_n)_{n=1}^N\}$, the visual-semantic embedding aims to map both images $\{(i_n)_{n=1}^N\}$ and sentences $\{(t_n)_{n=1}^N\}$ into a common space. In previous methods, for each image i_q, only the corresponding sentence t_q is regarded as relevant, and the others $\{t_p, (p \in \mathcal{N}^{-q})\}$ are all regarded as irrelevant, where $\mathcal{N}^{-q} = \{n | 1 \leq n \leq N, \text{and } n \neq q\}$. Thus, only the inequality $s(i_q, t_q) > s(i_q, t_p), (p \in \mathcal{N}^{-q})$ is enforced in previous methods.

In contrast, our approach will measure the semantic relevance degree between i_q and each sentence in $\{t_p, (p \in \mathcal{N}^{-q})\}$. Intuitively, the corresponding sentence t_q should have the highest relevance degree, while the others would have different degrees. Thus, in our coherent embedding space, the similarity of an image-sentence pair with higher relevance degree is desired to be greater than the similarity for a pair with lower degree.

To this end, we first define a continuous variable to measure the semantic relevance degree between images and sentences (in Sec.). Subsequently, to learn a coherent embedding space, we design a novel ladder loss to push different candidates away by distinct margins according to their relevance degree (in Sec.). At last, we propose the Coherent Score metric to properly measure whether the ranking order is aligned with their relevance degrees (in Sec.).

Our approach only relies on customized loss function and it has no restrictions on the image/sentence representation, so it is flexible to be incorporated into any neural network architecture.

Relevance Degree

In our approach, we need to measure the semantic relevance degree for image-sentence pairs. The ideal ground-truth for image-sentence pair is human annotation, but in fact it is infeasible to annotate such a multi-modal pairwise relevance dataset due to the combinatorial explosion in the number of possible pairs. On the other hand, the single-modal relevance measurement (*i.e.*, between sentences) is often much easier than the cross-modal one (*i.e.*, between sentences and images). For example, recently many newly proposed Natural Language Processing (NLP) models (Devlin et al. 2018; Peters et al. 2018; Liu et al. 2019) achieved very impressive results (Wang et al. 2018) on various NLP tasks. Specifically, on the sentence similarity task the BERT (Devlin et al. 2018) has nearly reached human performance. Compared to single-modal metric learning in image modality, the natural language similarity measure is more mature. Hence we cast the image-sentence relevance problem as a sentence-sentence relevance problem.

Intuitively, for an image i_q, the relevance degree of its corresponding sentence t_q is supposed to be the highest, and it is regarded as a reference when measuring the relevance degrees between i_q and other sentences. In other words, measuring the relevance degree between the image i_q and the

sentence t_p, $(p \in \mathcal{N})$ is cast as measuring the relevance degree (i.e. similarity) between the two sentences t_q and t_p, $(p \in \mathcal{N})$.

To this end, we employ the Bidirectional Encoder Representations Transformers (BERT) (Devlin et al. 2018). Specifically, the BERT model we used is fine-tuned on the Semantic Textual Similarity Benchmark (STS-B) dataset(Cer et al. 2017; Devlin et al. 2018). The Pearson correlation coefficient of our fine-tuned BERT on STS-B validation set is 0.88, which indicates good alignment between predictions and human perception. In short, the relevance degree between an image i_q and a sentence t_p is calculated as the similarity score between t_q and t_p with our fine-tuned BERT model:

$$R(i_q, t_p) = R(t_q, t_p) = \text{BERT}(t_q, t_p). \quad (1)$$

Ladder Loss Function

In this section, the conventional triplet loss is briefly overviewed, followed by our proposed ladder loss.

Triplet Loss   Let v_q be the visual representation of a query image i_q, and h_p indicates the representation of the sentence t_p. In the triplet loss formulation, for query image i_q, only its corresponding sentence t_q is regarded as the positive (*i.e.*, relevant) sample; while all other sentences $\{t_p, (p \in \mathcal{N}^{-q})\}$ are deemed negative (*i.e.*, irrelevant). Therefore, in the embedding space the similarity between v_q and h_q is encouraged to be greater than the similarity between v_q and $h_p, (p \in \mathcal{N}^{-q})$ by a margin α,

$$s(v_q, h_q) - s(v_q, h_p) > \alpha, (p \in \mathcal{N}^{-q}), \quad (2)$$

which can be transformed as the triplet loss function,

$$L_{tri}(q) = \sum_{p \in \mathcal{N}^{-q}} [\alpha - s(v_q, h_q) + s(v_q, h_p)]_+, \quad (3)$$

where $[x]_+$ indicates $\max\{0, x\}$. Considering the reflexive property of the query and candidate, the full triplet loss is

$$\mathcal{L}_{tri}(q) = \sum_{p \in N^{-q}} [\alpha - s(v_q, h_q) + s(v_q, h_p)]_+ \\ + \sum_{p \in N^{-q}} [\alpha - s(h_q, v_q) + s(h_q, v_p)]_+. \quad (4)$$

Ladder Loss   We first calculate the relevance degrees between image i_q and each sentence $t_p, (p \in \mathcal{N}^{-q})$. After that, these relevance degree values are divided into L levels with thresholds $\theta_l, (l = 1, 2, \ldots, L-1)$. As a result, the sentence index set $\mathcal{N}^{-q}$ is divided into L subsets $\mathcal{N}_1^{-q}, \mathcal{N}_2^{-q}, \ldots, \mathcal{N}_L^{-q}$, and sentences in $\mathcal{N}_l^{-q}$ are more relevant to the query than the sentences in $\mathcal{N}_{l+1}^{-q}$.

To learn a coherent embedding space, the more relevant sentences should be pulled closer to the query than the less relevant ones. To this end, we extend the single inequality Eq. (2) to an inequality chain,

$$\begin{aligned} s(v_q, h_q) - s(v_q, h_i) &> \alpha_1, (i \in \mathcal{N}_1^{-q}), \\ s(v_q, h_i) - s(v_q, h_j) &> \alpha_2, (i \in \mathcal{N}_1^{-q}, j \in \mathcal{N}_2^{-q}), \\ s(v_q, h_j) - s(v_q, h_k) &> \alpha_3, (j \in \mathcal{N}_2^{-q}, k \in \mathcal{N}_3^{-q}), \\ &\cdots, \end{aligned} \quad (5)$$
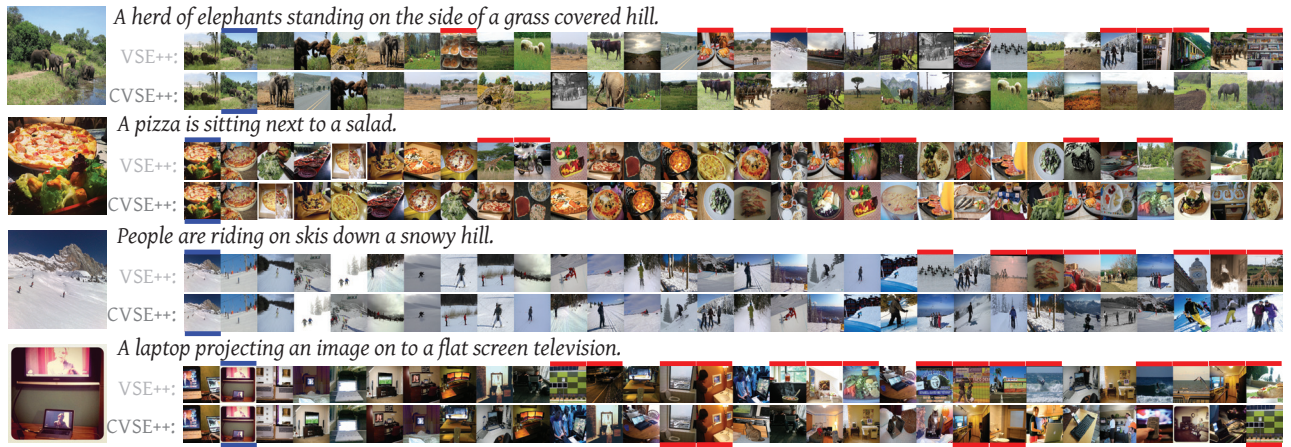
Figure 2: Comparison of the sentence-to-image top-30 retrieval results between VSE++ (baseline, 1st row) and CVSE++ (Ours, 2nd row). For each query sentence, the ground-truth image is shown on the left, the totally-relevant and totally-irrelevant retrieval results are marked by blue and red overlines/underlines, respectively. Despite that both methods retrieve the totally-relevant images at identical ranking positions, the baseline VSE++ method includes more totally-irrelevant images in the top-30 results; while our proposed CVSE++ method mitigates such problem.

where $\alpha_1, \ldots, \alpha_L$ are the margins between different non-overlapping sentence subsets.

In this way, the sentences with distinct relevance degrees are pushed away by distinct margins. For examples, for sentences in $\mathcal{N}_1^{-q}$, they are pushed away by margin α_1, and for sentences in $\mathcal{N}_2^{-q}$, they are pushed away by margin $\alpha_1 + \alpha_2$. Based on such inequality chain, we could define the ladder loss function. For simplicity, we just show the ladder loss with three-subset-partition (*i.e.*, $L = 3$) as an example,

$$L_{lad}(q) = \beta_1 L_{lad}^1(q) + \beta_2 L_{lad}^2(q) + \beta_3 L_{lad}^3(q), \quad (6)$$
$$L_{lad}^1(q) = \sum_{i \in \mathcal{N}_{1:L}^{-q}} [\alpha_1 - s(v_q, h_q) + s(v_q, h_i)]_+,$$
$$L_{lad}^2(q) = \sum_{i \in \mathcal{N}_1^{-q}, j \in \mathcal{N}_{2:L}^{-q}} [\alpha_2 - s(v_q, h_i) + s(v_q, h_j)]_+, (7)$$
$$L_{lad}^3(q) = \sum_{j \in \mathcal{N}_2^{-q}, k \in \mathcal{N}_{3:L}^{-q}} [\alpha_3 - s(v_q, h_j) + s(v_q, h_k)]_+,$$

where β_1, β_2 and β_3 are the weights between $L_{lad}^1(q)$, $L_{lad}^2(q)$ and $L_{lad}^3(q)$, respectively. $\mathcal{N}_{l:L}^{-q}$ indicates the union from $\mathcal{N}_l^{-q}$ to $\mathcal{N}_L^{-q}$.

As can be expected, the $L_{lad}^1(q)$ term alone is identical to the original triplet loss, *i.e.*, the ladder loss degenerates to the triplet loss if $\beta_2 = \beta_3 = 0$. Note that the dual problem of sentence as a query and images as candidates also exists. Similar to obtaining the full triplet loss Eq. (4), we can easily write the full ladder loss $\mathcal{L}_{lad}(q)$, which is omitted here.

Ladder Loss with Hard Contrastive Sampling For visual-semantic embedding, the hard negative sampling strategy (Simo-Serra et al. 2015; Wu et al. 2017) has been validated for inducing significant performance improvements, where selected hard samples (instead of all samples) are utilized for the loss computation. Inspired by (Wu et al. 2017; Faghri et al. 2018), we develop a similar strategy of selecting hard contrastive pairs for the ladder loss computation, which is termed **hard contrastive sampling (HC)**.

Taking the $L_{lad}^2(q)$ in Eq. (7) as an example, instead of conducting the sum over the sets $i \in \mathcal{N}_1^{-q}$ and $j \in \mathcal{N}_{2:L}^{-q}$, we sample one or several pairs (h_i, h_j) from $i \in \mathcal{N}_1^{-q}$ and $j \in \mathcal{N}_{2:L}^{-q}$. Our proposed HC sampling strategy involves choosing the h_j closest to the query in $\mathcal{N}_{2:L}^{-q}$, and the h_i furthest to the query in $\mathcal{N}_1^{-q}$ for the loss computation. Thus, the ladder loss part $L_{lad}^2(q)$ with hard contrastive sampling can be written as,

$$L_{lad-HC}^2(q) = [\alpha_1 - s(v_q, h_{i^*}) + s(v_q, h_{j^*})]_+,$$
$$j^* = \arg\max_{j \in \mathcal{N}_{2:L}^{-q}} s(v_q, h_j), \quad (8)$$
$$i^* = \arg\min_{i \in \mathcal{N}_1^{-q}} s(v_q, h_i),$$

where (i^*, j^*) is the index of the hardest contrastive pair (h_{i^*}, h_{j^*}). According to our empirical observation, this HC strategy not only reduces the complexity of loss computation, but also improves the overall performance.

Coherent Score

In previous methods, the most popular metric for visual-semantic embedding is R@K, which only accounts for the ranking position of the ground-truth candidates (*i.e.*, the totally-relevant candidates) while neglects others. Therefore, we propose a novel metric Coherent Score (CS) to properly measure the ranking order of all top-N candidates (including the ground-truth and other candidates).

The CS@K is defined to measure the alignment between the real ranking list $r_1, r_2, \ldots, r_K$ and its expected ranking list $e_1, e_2, \ldots, e_K$, where thee expected ranking list is decided according to their relevance degrees. We adopt Kendall's rank correlation coefficient τ, ($\tau \in [-1, 1]$) (Kendall 1945) as the criterion. Specifically, any pair of (r_i, e_i) and (r_j, e_j) where $i < j$ is defined to be concordant if both $r_i > r_j$ and $e_i > e_j$, or if both $r_i < r_j$ and

13053

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.

$e_i < e_j$. Conversely, it is defined to be discordant if the ranks for both elements mismatch. The Kendall's rank correlation τ depends on the number of concordant pairs and discordant pairs. When $\tau = 1$, the alignment is perfect, *i.e.* the two ranking lists are identical. Thus, a high CS@K score indicates the good quality and good user experience of the learnt embedding space and retrieval result in terms of coherence, and a model that achieves high CS@K score is expected to perform better in long-tail query challenges (Downey, Dumais, and Horvitz 2007) where a perfect match to the query does not necessarily exist in the database.

Experiments

Following related works, Flickr30K (Plummer et al. 2015) and MS-COCO (Lin et al. 2014; Chen et al. 2015) datasets are used in our experiments. The two datasets contain $31,000$ and $123,000$ images, respectively, and each image within them is annotated with 5 sentences using AMT. For Flickr30K, we use $1,000$ images for validation, $1,000$ for testing and the rest for training, which is consistent with (Faghri et al. 2018). For MS-COCO, we also follow (Faghri et al. 2018) and use $5,000$ images for both validation and testing. Meanwhile, the rest $30,504$ images in original validation set are used for training ($113,287$ training images in total) in our experiments following (Faghri et al. 2018). Our experimental settings follow that in VSE++ (Faghri et al. 2018), which is the state-of-the-art for visual-semantic embedding. Note, in terms of image-sentence cross modal retrieval, SCAN (Lee et al. 2018) achieves better performance, but it does not learn a joint embedding space for full sentences and full images, and suffers from combinatorial explosion in the number of sample pairs to be evaluated.

VGG-19 (Simonyan and Zisserman 2014) or ResNet-152 (He et al. 2016)-based image representation is used for our experiments (both pre-trained on ImageNet). Following common practice, we extract 4096 or 2048-dimensional feature vectors directly from the penultimate fully connected layer from these networks. We also adopt random cropping in data augmentation, where all images are first resized to 256×256 and randomly cropped 10 times at 224×224 resolution. For the sentence representation, we use a Gated Recurrent Unit (GRU), similar to the one used in (Faghri et al. 2018). The dimension of the GRU and the joint embedding space is set at $D = 1024$. The dimension of the word embeddings used as input to the GRU is set to 300.

Additionally, Adam solver is used for optimization, with the learning rate set at 2e-4 for 15 epochs, and then decayed to 2e-5 for another 15 epochs. We use a mini-batch of size 128 in all experiments in this paper. Our algorithm is implemented in PyTorch (Paszke et al. 2017).

Relevance Degree

The BERT inference is highly computational expensive (*e.g.*, a single NVIDIA Titan Xp GPU could compute similarity score for only approximately 65 sentence pairs per second). Therefore, it is computational infeasible to directly use Eq. (1) in practice due to combinatorial explosion of the number of sentence pairs.

In this paper, we mitigate the problem by introducing a coarse-to-fine mechanism. For each sentence pair we first employ conventional CBoW (Wang et al. 2018) method to coarsely measure their relevance degree. If the value is larger than a predefined threshold, Eq. (1) is used to refine their relevance degree calculation. The CBoW method first calculates each sentence's representation by averaging the GloVe (Pennington, Socher, and Manning 2014) word vectors for all tokens, and then computes the cosine similarity between their representations of each sentence pair. With this mechanism, the false-positive "relevant" pairs found by the CBoW method would be suppressed by BERT, while those important real relevant pairs would be assigned with more accurate relevance degrees. Thus, the speed of CBoW and the accuracy of BERT are combined properly. We empirically fix the predefined threshold at 0.8 for our experiments, as the mechanism achieves 0.79 in person correlation on STS-B.

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.

Results on MS-COCO

We compare VSE++ (re-implemented) and our Coherent Visual-Semantic Embedding (CVSE++) on the MS-COCO dataset, where VSE++ only focuses on the ranking position of the totally-relevant candidates while our approach cares about the ranking order of all Top-N candidates. The method of VSE++ (Faghri et al. 2018) is our baseline since it is the state-of-the-art approach for learning visual-semantic embedding. For fair comparison, we use both Recall@K (denoted as "R@K") and CS@K as metrics for evaluation, and also fine-tune (denoted by "FT") the CNNs following the baseline. In our approach, the hard contrastive sampling strategy is used. Experiments without the hard negative or hard contrastive sampling strategy are omitted because they perform much worse in terms of R@K, as reported in (Faghri et al. 2018).

In our approach, we need to determine the ladder number L in the loss function, which depends on how many top-ranked candidates (the value of N) we care about (*i.e.*, termed the scope-of-interest in this paper). With a small scope-of-interest, *e.g.*, top-100, only a few ladders are required, *e.g.*, $L = 2$; but with a larger scope-of-interest, *e.g.*, top-200, we will need more ladders, *e.g.*, $L = 3$, so that the low-level ladder, *e.g.*, $L_{lad}^2(q)$ in Eq. (6), is responsible for optimizing the ranking order of the very top candidates, *e.g.*, top-1 $\sim$ top-100; while the high-level ladder, *e.g.*, $L_{lad}^3(q)$ in Eq. (6), is responsible for optimizing the ranking order of subsequent candidates, *e.g.*, top-100 $\sim$ top-200.

A detailed discussion regarding the scope-of-interest and the choice of ladder number L will be provided in the next section. Practically, we limit our illustrated results to $L = 2$ both for computational savings and for the limited scope-of-interest from most human users. With ladder number L fixed at 2, parameters can be empirically determined by exploiting the validation set, *e.g.*, the threshold θ_1 for splitting $\mathcal{N}_1^{-q}$ and $\mathcal{N}_2^{-q}$ is fixed at 0.63, and the margins $\alpha_1 = 0.2$, $\alpha_2 = 0.01$, the loss weights $\beta_1 = 1$, $\beta_2 = 0.25$.

With our proposed CS@K metric, significantly larger K values are chosen than those (*e.g.*, $1, 5, 10$) in the classical R@K metric. For instance, we report the CS@100 and CS@1000 with 1000 test samples. Such choices of K allow more insights into both the local and global order-preserving effects in embedding space. In addition, the conventional R@K metrics are also included to measure the ranking performance of the totally-relevant candidates.

The experimental results on the MS-COCO dataset are presented in Tab. 1, where the proposed CVSE++ approaches evidently outperform their corresponding VSE++ counterparts in terms of CS@K, *e.g.*, from VSE++(Res152): 0.238 to CVSE++(Res152): 0.265 in terms of CS@100 for image→sentence retrieval with 1000 MS-COCO test samples. Moreover, the performance improvements are more significant with the larger scope-of-interest at CS@1000, *e.g.*, where "CVSE++ (Res152,FT)" achieves over 5-fold increase over "VSE++ (Res152,FT)" (from 0.071 to 0.446) in image→sentence retrieval. The result indicates that with our proposed ladder loss a coherent embedding space could be effectively learnt, which could produce significantly better ranking results especially in the global scope.

Simultaneously, a less expected phenomenon can be observed from Tab. 1: our proposed CVSE++ variants achieve roughly comparable or marginally better performance than their VSE++ counterparts in terms of R@K, *e.g.*, from VSE++(Res152): 63.2 to CVSE++(Res152): 66.7 in terms of R@1 for image→sentence retrieval with 1000 MS-COCO test samples. The overall improvement in R@K is insignificant because it completely neglects the ranking position of those non-ground-truth samples, and CVSE++ is not designed for improving the ranking for ground-truth. Based on these results, we speculate that the ladder loss appears to be beneficial (or at least not harmful) to the inference of totally-relevant candidates. Nevertheless, there are still hyper-parameters $(\beta_1, \beta_2, \cdots, \beta_L)$ controlling the balance between the totally-relevant and somewhat-relevant candidates, which will be further analyzed in the next section.

To provide some visual comparison between VSE++ and CVSE++, several sentences are randomly sampled from the validation set as queries, and their corresponding retrievals are illustrated in Fig. 2 (sentence→image). Evidently, our CSVE++ could put more somewhat-relevant candidates and reduce the number of totally-irrelevant candidates on the top-N retrieval list and enhance user experience.

Results on Flickr30K

Our approach is also evaluated on the Flikr30K dataset and compared with the baseline VSE++ variants, as shown in Tab. 2. The hyper-parameter settings are identical to that in Tab. 1 with MS-COO (1000 Test Samples). As expected, these experimental results demonstrate similar performance improvements both in terms of CS@K and R@K by our proposed CVSE++ variants.

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).

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)

Parameter Sensitivity Analysis

In this section, parameter sensitivity analysis is carried out on two groups of hyper-parameters, *i.e.*, the balancing parameter $\beta_1, \beta_2, \cdots, \beta_L$ in Eq. (6) and the ladder number L.

Balancing Totally Relevant and Others

In Eq. (6), the weights between the ranking position optimization of totally-relevant candidates and other candidates in the ladder loss are controlled by the hyper-parameters $\beta_1, \beta_2, \cdots, \beta_L$. With $\beta_2 = \cdots = \beta_L = 0$, the ladder loss degenerates to the triplet loss, and all emphasis is put on the totally-relevant ones. Conversely, relatively larger $\beta_2, \cdots, \beta_L$ values put more emphasis on the somewhat-relevant candidates.

With other parameters fixed (L fixed at 2, β_1 fixed at 1), parameter sensitivity analysis is carried out on β_2 only. From Tab. 3, we can see that CS@K metrics improve with larger β_2, but R@K metrics degrade when β_2 is close to 1.0. Based on the three β_2 settings in Tab. 3, we speculate that CS@K and R@K metrics would not necessarily peak simultaneously at the same β_2 value. We also observe that with excessively large β_2 values, the R@K metrics drop dramatically. Generally, the ranking orders of the totally-relevant candidates often catch user's attention and they should be optimized with high priority. Therefore, we select $\beta_2 = 0.25$ in all our other experiments to strike a balance because of R@K and CS@K performance.

The Scope-of-interest for Ladder Loss

Our approach focuses on improving the ranking order of all top-N retrieved results (instead of just the totally-relevant ones). Thus, there is an important parameter, *i.e.*, the scope-of-interest N or the size of the desired retrieval list. If the retrieval system user only cares about a few top-ranked results (*e.g.*, top-100), two ladders (*e.g.*, $L = 2$) are practically sufficient; If a larger scope-of-interest (*e.g.*, top-200) is required, more ladders are probably needed in the ladder loss. For example, with $L = 3$, the low-level ladder $L_{lad}^2(q)$ is responsible for the optimization of the ranking order of very top candidates, *e.g.*, from top-1 ∼ top-100; while the high-level ladder $L_{lad}^3(q)$ is responsible for the optimization of the ranking order of subsequent candidates, *e.g.*, from top-100 ∼ top-200. Inevitably, larger ladder number results in

higher computational complexity. Therefore, a compromise between the scope-of-interest and the computational complexity needs to be reached.

For the sensitivity analysis of ladder number $L = 1, 2, 3$, we evaluate our CVSE++ (Res152) approach by comparing top-100, top-200 and top-1000 results, which are measured by CS@100, CS@200 and CS@1000, respectively. Other parameters $\theta_2, \alpha_3, \beta_3$ are empirically fixed at 0.56, 0.01, 0.125, respectively. The experimental results are summarized in Tab. 4. With small scope-of-interest $N = 100$, we find that two ladder $L = 2$ is effective to optimize the CS@100 metric, a third ladder only incurs marginal improvements. However, with larger scope-of-interest, *e.g.*, top-200, the CS@200 can be further improved by adding one more ladder, *i.e.*, $L = 3$.

Apart from that, a notable side effect with too many ladders (*e.g.* 5) can be observed, the R@K performance drops evidently. We speculate that with more ladders, the ladder loss is likely to be dominated by high-level ladder terms and leads to some difficulties in optimization of the low-level ladder term. This result indicates that the choice of L should be proportional to the scope-of-interest, *i.e.*, more ladders for larger scope-of-interest and vice versa.

Conclusion

In this paper, relevance between queries and candidates are formulated as a continuous variable instead of a binary one, and a new ladder loss is proposed to push different candidates away by distinct margins. As a result, we could learn a coherent visual-semantic space where both the totally-relevant and the somewhat-relevant candidates can be retrieved and ranked in a proper order.

In particular, our ladder loss improves the ranking quality of all top-N results without degrading the ranking positions of the ground-truth candidates. Besides, the scope-of-interest is flexible by adjusting the number of ladders. Extensive experiments on multiple datasets validate the efficacy of our proposed method, and our approach achieves the state-of-the-art performance in terms of both CS@K and R@K. For future work, we plan to extend the ladder loss-based embedding to other metric learning applications.

Acknowledgements

References

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *ICML*, 1247–1255.

Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; and Specia, L. 2017. SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation. *ArXiv e-prints*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; and Zitnick, C. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Downey, D.; Dumais, S.; and Horvitz, E. 2007. Heads and tails: studies of web search with common and rare queries. In *ACM SIGIR*, 847–848.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.

Frome, A.; Corrado, G.; Shlens, J.; Bengio, S.; Dean, J.; and Ranzato, T. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*.

Gong, Y.; Ke, Q.; Isard, M.; and Lazebnik, S. 2014a. A multiview embedding space for modeling internet images, tags, and their semantics. *IJCV* 106(2):210–233.

Gong, Y.; Wang, L.; Hodosh, M.; Hockenmaier, J.; and Lazebnik, S. 2014b. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, 529–545.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 1735–1742.

Hardoon, D. R.; Szedmak, S.; and Shawe-Taylor, J. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation* 16(12):2639–2664.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Huang, C.; Loy, C. C.; and Tang, X. 2016. Local similarity-aware deep feature embedding. In *NIPS*, 1262–1270.

Ji, X.; Wang, W.; Zhang, M.; and Yang, Y. 2017. Cross-domain image retrieval with attention modeling. In *ACM MM*, 1654–1662.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137.

Karpathy, A.; Joulin, A.; and Fei-Fei, L. 2014. Deep fragment embeddings for bidirectional image-sentence mapping. In *NIPS*.

Kendall, M. G. 1945. The treatment of ties in ranking problems. *Biometrika* 33(3):239–251.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *NIPS*.

Klein, B.; Lev, G.; Sadeh, G.; and Wolf, L. 2014. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 201–216.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Liu, X.; He, P.; Chen, W.; and Gao, J. 2019. Multi-task deep neural networks for natural language understanding. *CoRR* abs/1901.11504.

Malinowski, M.; Rohrbach, M.; and Fritz, M. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 1–9.

Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; and Hua, G. 2017. Hierarchical multimodal lstm for dense visual-semantic embedding. In *ICCV*, 1899–1907.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *CVPR*, 4004–4012.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL*.

Plummer, B.; Wang, L.; Cervantes, C.; Caicedo, J.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *ICCV*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; and Moreno-Noguer, F. 2015. Discriminative learning of deep convolutional feature point descriptors. In *ICCV*, 118–126.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.

Socher, R.; Le, Q.; Manning, C.; and Ng, A. 2014. Grounded compositional semantics for finding and describing images with sentences. *TACL*.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 1857–1865.

Ustinova, E., and Lempitsky, V. 2016. Learning deep embeddings with histogram loss. In *NIPS*, 4170–4178.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning deep structure-preserving image-text embeddings. *CVPR*.

Wu, C.-Y.; Manmatha, R.; Smola, A. J.; and Krähenbühl, P. 2017. Sampling matters in deep embedding learning. In *ICCV*.

Yan, F., and Mikolajczyk, K. 2015. Deep correlation for matching images and text. In *CVPR*, 3441–3450.

Yuan, Y.; Yang, K.; and Zhang, C. 2017. Hard-aware deeply cascaded embedding. In *ICCV*, 814–823.