This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3139210, IEEE Transactions on Multimedia

Adaptive Ladder Loss for Learning Coherent Visual-Semantic Embedding

Le Wang, *Senior Member, IEEE,* Mo Zhou, *Student Member, IEEE,* Zhenxing Niu, *Member, IEEE,* Qilin Zhang, *Member, IEEE,* Nanning Zheng, *Fellow, IEEE*

***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. To adapt to the varying mini-batch statistics and improve the efficiency of the ladder loss, we also propose a Silhouette score-based method to adaptively decide the ladder level and hence the underlying inequality chain. 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.**

***Index Terms*—Coherent Visual-Semantic Embedding, Adaptive Ladder Loss, Hard-Contrastive Sampling, Coherent Score.**

I. 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 [1]. Recently, the advances in deep learning have made significant progress on visual-semantic embedding [2], [3], [4], [5]. 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 [6], [7].

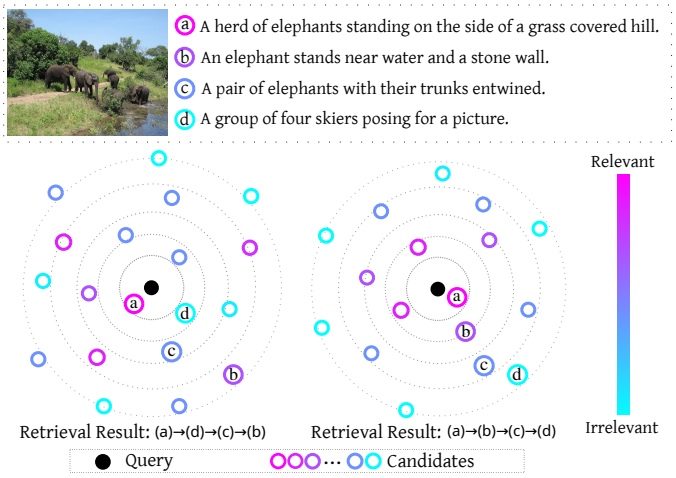


Fig. 1. Comparison between the incoherent (left) and coherent (right) visual-semantic embedding space. Existing methods (left) pull the totally-relevant sentence (a) close to the query image, while pushing away all other sentences (b, c, and d) equally. Therefore, the relative proximity of (b, c, and 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 [8], [6] 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 for coherent results.

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) [2], [3], [5] 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

Manuscript received April 09, 2021; revised XXX XX, 2021; accepted XXX XX, 2021. This work was supported partly by National Key R&D Program of China Grant 2018AAA0101400, NSFC Grants 62088102, 61773312, and 61976171, and Fundamental Research Funds for the Central Universities under Grant XTR042021005. *(Corresponding author: Le Wang.)*

L. Wang, M. Zhou, and N. Zheng are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: {lewang, nnzheng}@mail.xjtu.edu.cn, cdluminate@gmail.com).

Z. Niu is with Alibaba Group, Hangzhou, Zhejiang 311121, China (e-mail:niuzhenxing@gmail.com).

Q. Zhang is with ABB Corporate Research Center, Raleigh, NC 27606, USA (e-mail: samqzhang@gmail.com). This research work was carried out before his joining of ABB.

(known as the long-tail query challenge [9]). Given a query from the user, when an 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 Figure 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 $\mathcal{I}_q$ and its totally-relevant sentence $\mathcal{T}_q$ in the conventional triplet loss [5] is encouraged to be greater than the similarity between the $\mathcal{I}_q$ and other sentences $\mathcal{T}_p$. Likewise, with the ladder loss formulation, we consider the relevance degrees of all sentences, and extend the inequality $s(\mathcal{I}_q, \mathcal{T}_q) > s(\mathcal{I}_q, \mathcal{T}_p)$ to an inequality chain, *i.e.*, $s(\mathcal{I}_q, \mathcal{T}_q) > s(\mathcal{I}_q, \mathcal{T}_{p_1}) > s(\mathcal{I}_q, \mathcal{T}_{p_2}) > \cdots > s(\mathcal{I}_q, \mathcal{T}_{p_L})$, where $\mathcal{T}_{p_l}$ is more relevant to $\mathcal{I}_q$ than $\mathcal{T}_{p_{l+1}}$, 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.

Albeit the underlying inequality chains could be decided after grouping samples with manually specified relevance degree thresholds, the varying mini-batch statistics (*i.e.*, the distribution of the relevance degree values) during the training process may result in inefficiency for the ladder loss. For instance, when all the relevance degrees of the *not* totally-relevant samples within the mini-batch are less than the given threshold, ladder loss will gracefully degenerate into the fundamental triplet loss and discard any relevance information. Namely, fixed thresholds are not flexible to better exploit the relevance information from the "somewhat relevant" samples. To this end, we extend the ladder loss with a Silhouette score-based automatic K-Means clustering method for adaptively selecting ladder levels to dynamically form the inequalities for each mini-batch. With this method incorporated, adaptive ladder loss can better exploit relevance degree information and further boost the coherence of embedding space.

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 proposed by extending the inequality in the triplet loss to an inequality chain, so that candidates with different degrees will be treated differently.
3) To improve ladder loss efficiency for a more coherent embedding space, we propose a Silhouette score-based method to adaptively decide the inequality chains during the training process due to varying mini-batch statistics.
4) 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.

This paper is an extension of our previous conference paper [10] with an adaptive extension of the previous approach, more technical details, and improved readability. In brief, the changes include (1) a new adaptive version of the ladder loss which can adaptively decide the underlying inequality chain; (2) More figures, illustrations and details on the proposed loss function, the adaptive selection of ladder levels, and the coherent score; (3) Replacement of the fine-tuned BERT with Sentence-BERT [11] for relevance degree; (4) Fully updated experimental results including new parameter searching experiments for the adaptive version of ladder loss; (5) More visualizations of retrieval results on coherent visual-semantic embedding; (6) Reorganization of the mathematical symbols in order to avoid ambiguity; (7) Introduction of more recent related works.

The remainder of this paper is organized as follows. Section II briefly reviews the related works on visual-semantic embedding and metric learning. The problem formulation, relevance degree, ladder loss and its adaptive extension, and the coherence score will be introduced in Section III. Extensive experiments with detailed discussions are presented in Section IV and Section V. In the last section, the whole paper is concluded.

II. Related Work

A. Visual-Semantic Embedding

Visual-semantic embedding, as a kind of multi-modal joint embedding, enables a wide range of tasks such as image-caption retrieval [4], [2], [5], image captioning, and visual question-answering [12]. Generally, the methods of visual-semantic embedding could be divided into two categories. The first category is based on Canonical Correlation Analysis (CCA) [13], [14], [15], [16] 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 [17], [18].

The second category involves metric learning-based embedding space learning [19], [20], [5]. DeViSE [19], [21] learns linear transformations of visual and textual features

to the common space. After that, Deep Structure-Preserving (DeepSP) [20] is proposed for image-text embedding, which combines cross-view ranking constraints with within-view neighborhood structure preservation. In [22], 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.* [5] incorporate hard negatives in the ranking loss function, which yields significant gains in retrieval performance. To further narrow the heterogeneity gap between different modalities, Lin [23] explores the similarity between inter-modal instances. Yu [24] proposes a plug-and-play module to capture rich visual semantics and help to enhance the visual representation for cross-modal analysis. Ma [25] propose to leverage multi-level correlation information for hashing. Compared to CCA-based methods, metric learning-based methods scale better to large dataset with stochastic optimization in training.

B. Deep Metric learning and Image Retrieval

Deep metric learning has many other applications such as face recognition [6], person re-identification [26] and fine-grained recognition [27], [28], [29]. The loss function design and sample mining strategy in metric learning could be subtle but important problems [30]. For example, the contrastive loss [8] 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 [6], 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 [31] and Histogram Loss [32] use quadruplets. Beyond that, the n-pair loss [7] and Lifted Structure [27] define constraints on all images in a batch.

Recently, Wang *et al.* [33] provide a multi-similarity loss with general pair weighting for understanding the recent pair-based loss functions. Zhou *et al.* [34] propose an adversarial-example-augmented triplet loss for enhancing image retrieval robustness against input image perturbations. Wang *et al.* [35] present a ranking-motivated structured loss that exploits all instances in the gallery. Xuan *et al.* [36] characterize the space of triplets and analyze the impact of hard negative examples to the triplet loss. Jun *et al.* [37] propose a framework to exploit multiple global descriptors to get an ensemble effect for image retrieval. Min *et al.* [38] propose a two-stage triplet network using auxiliary joint learning of classification and region-level supervision for image retrieval.

However, all the aforementioned methods formulate the relevance as a binary variable. Thus, our ladder loss could be used to boost those methods. Besides, since diversity is also an important factor for both image retrieval and text retrieval results [39], we argue that a coherent embedding is also beneficial for retrieval diversity, because less irrelevant and more relevant retrieval results will appear in the top.

III. Our Approach

Given a set of image-sentence pairs $\mathcal{D} = \{(\mathcal{I}_i, \mathcal{T}_i)_{i=1}^{\mathsf{N}}\}$, the visual-semantic embedding aims to map both images $\{(\mathcal{I}_i)_{i=1}^{\mathsf{N}}\}$ and sentences $\{(\mathcal{T}_i)_{i=1}^{\mathsf{N}}\}$ into a common space. In previous methods, for each image $\mathcal{I}_q$, only the corresponding sentence $\mathcal{T}_q$ is regarded as relevant, and the others $\{\mathcal{T}_p, (p \in \mathcal{N}^{-q})\}$ are all regarded as irrelevant, where $\mathcal{N}^{-q} = \{i | 1 \leq i \leq \mathsf{N}, \text{and } i \neq q\}$ is the index set of the not totally-relevant samples. Thus, only the inequality $s(\mathcal{I}_q, \mathcal{T}_q) > s(\mathcal{I}_q, \mathcal{T}_p), (p \in \mathcal{N}^{-q})$ is enforced in previous methods.

In contrast, our approach will measure the semantic relevance degree between $\mathcal{I}_q$ and each sentence in $\{\mathcal{T}_p, (p \in \mathcal{N}^{-q})\}$. Intuitively, the corresponding sentence $\mathcal{T}_q$ 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. 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.

To this end, we first define a continuous variable to measure the semantic relevance degree between images and sentences (in Section III-A). 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 Section III-B). training mini-batch statistics, we present an adaptive ladder loss in Section III-C. At last, we propose the Coherent Score metric to properly measure whether the ranking order is aligned with their relevance degrees (in Section III-D).

A. 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 [40], [41], [42] achieved very impressive results [43] on various NLP tasks. Specifically, on the sentence similarity task the BERT [40] 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 $\mathcal{I}_q$, the relevance degree of its corresponding sentence $\mathcal{T}_q$ is supposed to be the highest, and it is regarded as a reference when measuring the relevance degrees between $\mathcal{I}_q$ and other sentences. In other words, measuring the relevance degree between the image $\mathcal{I}_q$ and the sentence $\mathcal{T}_p,\ (p \in \mathcal{N})$ is cast as measuring the relevance degree (*i.e.*, similarity) between the two sentences $\mathcal{T}_q$ and $\mathcal{T}_p,\ (p \in \mathcal{N})$. Similarly, if the query is a sentence $\mathcal{T}_q$, the relevance degree between $\mathcal{T}_q$ and image $\mathcal{I}_p,\ (p \in \mathcal{N})$ can be measured by the relevance degree between the two sentences $\mathcal{T}_q$ and $\mathcal{T}_p, (p \in \mathcal{N})$, where the $\mathcal{T}_p,\ (p \in \mathcal{N})$ is the corresponding ground-truth sentence of the image $\mathcal{I}_p, (p \in \mathcal{N})$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3139210, IEEE Transactions on Multimedia

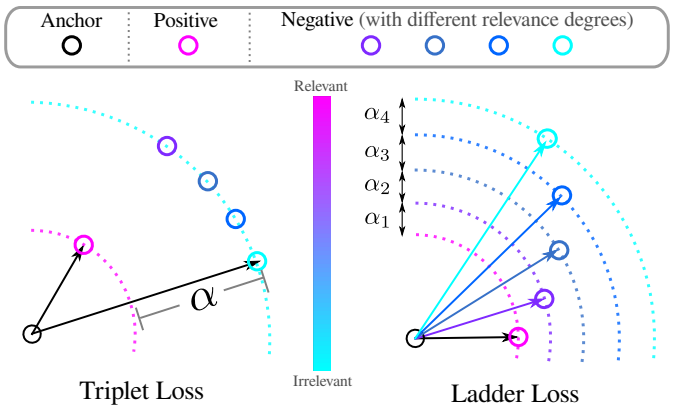


Fig. 2. Illustration of difference between Triplet Loss and our Ladder Loss. The triplet loss treats the relevance between query ("Anchor") and candidates ("Positive" and "Negative" examples with different relevance degrees to the query) in a bipolar way – either relevant or irrelevant, neglecting their different proximity to the query. In contrast, our proposed ladder loss treats the samples differently according to their relevance degrees. Ladder loss can aid the learning of a coherent visual-semantic embedding space.

To this end, we can employ the Bidirectional Encoder Representations Transformers (BERT) [40] fine-tuned on the Semantic Textual Similarity Benchmark (STS-B) dataset [44], [40] which achieves a Pearson correlation coefficient of 0.88 on the validation set of STS-B, indicating a good alignment between predictions and human perception. However, according to Li *et al.* [45], the semantic information in the BERT embeddings is not fully explored. Moreover, BERT requires that both sentences for similarity calculation are fed into the network, and result in a massive computational overhead [11], which makes it unsuitable for our task that needs to calculate sentence similarity intensively. To solve this problem, Reimers *et al.* [11] propose Sentence-BERT, a modification of BERT to derive semantically meaningful sentence embeddings that can be compared using cosine similarity.

In short, the relevance degree between an image $\mathcal{I}_q$ and a sentence $\mathcal{T}_p$ is calculated as the similarity score between $\mathcal{T}_q$ and $\mathcal{T}_p$ with Sentence-BERT [11]:

$$R(\mathcal{I}_q, \mathcal{T}_p) = R(\mathcal{T}_q, \mathcal{T}_p) = \text{Sentence-BERT}(\mathcal{T}_q, \mathcal{T}_p). \quad (1)$$

Ideally, a well-trained coherent visual semantic embedding space should be able to return retrieval results that are coherent to such relevance degree.

B. Ladder Loss Function

In this section, the conventional triplet loss is briefly overviewed, followed by our proposed ladder loss. A diagram illustrating the difference between the existing triplet loss and the proposed ladder loss is presented in Figure 2.

1) Triplet Loss: Let v_q be the visual representation of a query image $\mathcal{I}_q$, and h_p indicates the representation of the sentence $\mathcal{T}_p$. In the triplet loss formulation, for query image $\mathcal{I}_q$, only its corresponding sentence $\mathcal{T}_q$ is regarded as the positive (*i.e.*, relevant) sample; while all other sentences $\{\mathcal{T}_p, (p \in \mathcal{N}^{-q})\}$ 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,

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

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

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

2) Ladder Loss: We first calculate the relevance degrees between image $\mathcal{I}_q$ and each sentence $\mathcal{T}_p, (p \in \mathcal{N}^{-q})$. 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,

$$\begin{aligned}&s(v_q, h_q) - s(v_q, h_i) > \alpha_1, (i \in \mathcal{N}_1^{-q}), \\ &s(v_q, h_i) - s(v_q, h_j) > \alpha_2, (i \in \mathcal{N}_1^{-q}, j \in \mathcal{N}_2^{-q}), \\ &s(v_q, h_j) - s(v_q, h_k) > \alpha_3, (j \in \mathcal{N}_2^{-q}, k \in \mathcal{N}_3^{-q}), \\ &\cdots,\end{aligned} \quad (5)$$

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,

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

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

As can be expected, the $\mathcal{L}_{\text{lad}}^1(q)$ 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 $\mathcal{L}_{\text{lad}}(q)$, which is omitted here for sake of brevity. In all experiments of this paper, we always use

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3139210, IEEE Transactions on Multimedia

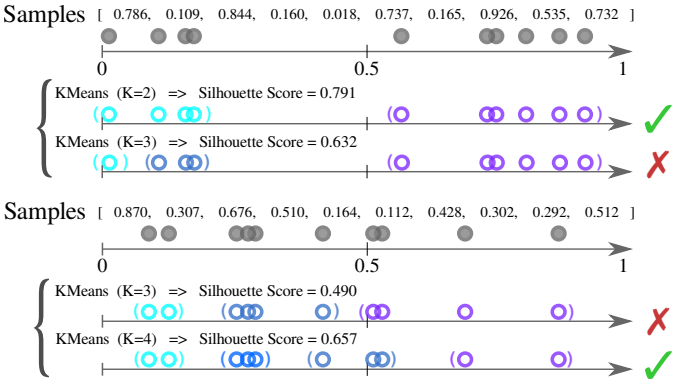


Fig. 3. Demonstration of adaptive ladder level selection based on K-Means clustering and Silhouette score. Given a batch of 10 samples and their relevance degrees to the query, for the upper example the mean Silhouette score for $k = 2$ is 0.791, while that for $k = 3$ is 0.632 ($2 \leq k \leq 5$; cases for $k = 4$ and $k = 5$ are omitted for brevity). Since the maximum Silhouette score corresponds to $k = 2$, the number of ladder levels is set as 2 for the current batch of samples, and the clusters for $k = 2$ is directly used as the grouping result, *i.e.* $\mathcal{N}_i^{-q}, i \in \{1, 2, \ldots, k\}$. Similarly, in the lower example the mean Silhouette score for $k = 4$ is the maximum for $2 \leq k \leq 5$ (cases for $k = 2$ and $k = 5$ are omitted for brevity). Best viewed in color.

the full (*i.e.*, "cycle-consistent") ladder loss, which contains both the image-to-sentence and sentence-to-image loss terms.

3) Hard Contrastive Sampling: For visual-semantic embedding, the hard negative sampling strategy [46], [28] has been validated for inducing significant performance improvements, where selected hard samples (instead of all samples) are utilized for the loss computation. Inspired by [28], [5], we develop a similar strategy of selecting hard contrastive pairs for the ladder loss computation, which is termed **hard contrastive sampling (HC)**.

Taking the $\mathcal{L}^2_{\text{lad}}(q)$ 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 $\mathcal{L}^2_{\text{lad}}(q)$ with hard contrastive sampling can be written as,

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

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.

C. Adaptive Ladder Loss

Since the proposed ladder loss function is based on a series of inequalities (*i.e.* Eq.5) across different sample sets, the result of sample grouping may greatly impact the efficacy of the loss function. Intuitively, the number L and the corresponding thresholds θ_l, $(l = 1, 2, \ldots, L-1)$ can be manually set, but due to the randomness of mini-batch sampling during the training process (*i.e.*, the relevance degree distribution in a training mini-batch is volatile), manually fixed thresholds may render inefficiency for learning a coherent embedding. For instance, sometimes the relevance degrees of all samples within a batch are smaller than the given thresholds, then ladder loss will simply degenerate into the triplet loss, due to lack of training samples in different ladder levels. Hence, manually tuning the thresholds is difficult, and the advantages of the proposed ladder loss would be weakened by fixed thresholds.

To this end, we propose an adaptive method for automatically grouping the non-positive samples $\mathcal{N}^{-q}$ into variable number of non-overlapping subsets $\mathcal{N}_1^{-q}, \mathcal{N}_2^{-q}, \ldots, \mathcal{N}_{L^*}^{-q}$ ($L_{\text{MIN}} \leq L^* \leq L_{\text{MAX}}$) beside grouping with manually set thresholds. We cast this problem into an automatic clustering problem for relevance degree values $R(\mathcal{I}_q, \mathcal{T}_i)$, $i \in \mathcal{N}^{-q}$. For brevity, we denote $R(\mathcal{I}_q, \mathcal{T}_i)$ as $R_{q,i}$. Then the relevance degree values can be clustered into L^* ($L_{\text{MIN}} \leq L^* \leq L_{\text{MAX}}$) sets according to the mean Silhouette value over all relevance degrees of samples in $\mathcal{N}^{-q}$, *i.e.*,

$$L^* = \arg\max_{L_{\text{MIN}} \leq k \leq L_{\text{MAX}}} \frac{1}{|\mathcal{N}^{-q}|} \sum_{i \in \mathcal{N}^{-q}} \text{Sil}\big(R_{q,i}; \{C_\xi\}_{\xi=1}^k\big), \tag{9}$$

where $\text{Sil}(\cdot)$ computes the Silhouette value, $\{C_\xi\}_{\xi=1}^k$ is the k sample clusters from the k-Means clustering algorithm. In practice, the most appropriate L^* can be found by evaluating the mean Silhouette values based on clustering results for each possible k. Thus, we only need to define the minimal and maximal ladder numbers L_{MIN} and L_{MAX}, and the problems of manually set thresholds are addressed.

In particular, we describe the Silhouette value $\text{Sil}(\cdot)$ [47] in detail. Given a set of relevance degree values that are clustered with K-Means algorithm into k clusters. Firstly, the mean intra-cluster distance $a(R_{q,i})$ between relevance degree $R_{q,i}$ and all other data points in the same cluster is defined as follows:

$$a(R_{q,i}) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} |R_{q,i} - R_{q,j}|, \tag{10}$$

whose value reflects how well $R_{q,i}$ is assigned to its cluster (the smaller the better). Then, the smallest mean distance of $R_{q,i}$ to all data points in any other cluster (of which $R_{q,i}$ is not included) is defined as:

$$b(R_{q,i}) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} |R_{q,i} - R_{q,j}|. \tag{11}$$

Finally, the silhouette value of a data point $R_{q,i}$ is defined as

$$\text{Sil}(R_{q,i}) = \frac{b(R_{q,i}) - a(R_{q,i})}{\max\{a(R_{q,i}), b(R_{q,i})\}} \tag{12}$$

when $|C_i| > 1$, or $\text{Sil}(R_{q,i}) = 0$ when $|C_i| = 1$. It is clear that $-1 < \text{Sil}(R_{q,i}) < +1$, and a large Silhouette value is preferred for a proper clustering result.

Thus, by using Silhouette value as a criterion, the ladder levels can be adaptively decided based on the varying mini-batch relevance degree statistics, no longer relying on manually selected thresholds to group the samples. The inequality chains are then directly obtained from the clustering results $\{C_\xi\}_{\xi=1}^{L^*}$. Namely, the adaptive sentence index sets $\mathcal{N}_\xi^{-q}$ correspond to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3139210, IEEE Transactions on Multimedia

Fig. 4. 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. Figure best viewed in color.

C_ξ. After incorporating this method into the proposed ladder loss, we obtain the adaptive version of ladder loss.

D. 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, $(\tau \in [-1, 1])$ [48] 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 $e_i < e_j$. Conversely, it is defined to be discordant if the ranks for both elements mismatch. When $r_i = r_j$ or $e_i = e_j$, the pair is defined to be tied. The Kendall's rank correlation τ depends on the number of concordant pairs and discordant pairs, and it is defined as

$$\tau = \frac{P - Q}{\sqrt{(P + Q + T)(P + Q + U)}}, \tag{13}$$

where P is the number of concordant pairs, Q the number of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3139210, IEEE Transactions on Multimedia

Fig. 5. Comparison of image-to-sentence top-30 retrieval results between VSE++ (baseline) and CVSE++ (ours). Each query image is shown on the left; and the top-30 retrieval sentences are listed on the right. Totally-relevant and totally-irrelevant sentences are highlighted in blue and red color, respectively. It is clear that both methods retrieve the totally-relevant sentences at high ranking positions, but the baseline VSE++ method includes much more totally-irrelevant sentences in the top-30 results; while our proposed CVSE++ method mitigates such problem. Figure best viewed in color.

discordant pairs, T, U the numbers of ties only in $r_1, r_2, \ldots, r_K$ or $e_1, e_2, \ldots, e_K$, respectively.

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 [9] where a perfect match to the query does not necessarily exist in the database.

It is noted that, although the two metrics, *i.e.*, CS@K and R@K peak at the same time in the ideal case, they are reflecting different aspects of the ranking result, and are not positively correlated. The R@K is purely based on the ranking position of the ground-truth candidates, while neglecting the ranking order of all other candidates (described in the third paragraph in Section I). Thus, the R@K is invariant to the order among non-ground-truth samples. Besides, the CS@K measures the alignment between the real ranking order and the expected ranking order. The position of the ground-truth candidate is merely a small fraction of the ranking order which decides the CS@K value, and the CS@K can be invariant to the ground-truth candidate in some cases.

Such invariance can be demonstrated with examples. Assume that we have a set of sentences $\{h_1, h_2, h_3, h_4, h_5\}$. Given an image embedding v_q, the ideal ranking order among the sentences is "h_1, h_2, h_3, h_4, h_5", where both CS@5$=$ 1.0 and R@1$=$ 100 reach the maximum value. When the real ranking list is "h_1, h_5, h_4, h_3, h_2" or "h_1, h_2, h_3, h_4, h_5", the corresponding (R@1, CS@5) values are (100, -0.2) or (100, 1.0), respectively. The previous ranking list is badly ordered, while the latter one is coherent, but both lists reach the highest R@K value. When the real ranking list is "h_2, h_1, h_3, h_4, h_5" or "h_1, h_3, h_2, h_4, h_5", the corresponding (R@1, CS@5) values are (0, 0.8) or (100, 0.8), respectively. Both lists are somewhat coherent, only containing one reversed pair compared to the ideal order, and hence achieve the same CS@5. However, their R@1 varies. In brief, the two metrics are not positively correlated, and are revealing different aspects of the ranking result.

TABLE I
COMPARISON BETWEEN VSE++ AND CVSE++ IN TERMS OF CS@K AND R@K ON MS-COCO DATASET.

MS-COCO Dataset (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)	0.264	0.073	4.3	56.0	84.0	92.0	0.268	0.077	10.1	42.5	78.0	88.0
CVSE++ (VGG19, [10])	0.281	0.175	3.7	56.5	85.1	92.9	0.273	0.178	7.3	43.0	78.8	88.7
CVSE++ (VGG19)	0.299	0.234	3.9	57.0	84.6	92.7	0.291	0.232	9.3	42.5	76.8	87.3
CVSE++ (VGG19, Auto)	0.300	0.208	4.2	56.8	83.3	93.1	0.293	0.200	9.5	41.5	77.6	87.5
VSE++ (VGG19, FT)	0.285	0.104	2.9	61.6	89.5	95.5	0.284	0.108	6.7	50.1	84.1	92.0
CVSE++ (VGG19, FT, [10])	0.313	0.206	2.8	63.1	89.5	96.3	0.293	0.214	5.3	50.3	84.6	92.5
CVSE++ (VGG19, FT)	0.316	0.292	3.0	63.6	90.0	96.2	0.295	0.283	6.3	50.0	83.7	92.0
CVSE++ (VGG19, FT, Auto)	0.331	0.283	3.1	63.2	89.0	95.5	0.313	0.270	6.2	49.5	84.8	92.3
VSE++ (Res152)	0.264	0.107	3.0	63.4	87.7	95.6	0.263	0.112	7.7	47.6	80.3	90.1
CVSE++ (Res152, [10])	0.290	0.192	2.9	66.5	89.0	95.2	0.278	0.196	6.0	48.6	80.5	90.5
CVSE++ (Res152)	0.301	0.265	3.1	65.2	88.6	95.0	0.292	0.266	6.9	48.3	79.8	89.3
CVSE++ (Res152, Auto)	0.305	0.236	3.1	64.7	88.1	95.2	0.295	0.229	7.1	47.2	79.1	89.5
VSE++ (Res152, FT)	0.263	0.099	2.4	67.8	91.1	96.5	0.262	0.103	6.4	53.8	85.1	92.1
CVSE++ (Res152, FT, [10])	0.308	0.205	2.4	69.3	92.5	97.0	0.292	0.219	4.4	55.4	86.3	93.8
CVSE++ (Res152, FT)	0.310	0.303	2.3	68.5	93.0	97.1	0.292	0.295	5.4	54.4	84.9	92.6
CVSE++ (Res152, FT, Auto)	0.313	0.292	2.5	68.0	91.7	96.0	0.297	0.292	5.4	53.6	85.7	93.1
MS-COCO Dataset (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)	0.254	0.109	11.0	36.3	65.9	78.6	0.235	0.113	29.6	25.5	53.5	66.5
CVSE++ (Res152, [10])	0.278	0.194	10.5	39.3	68.0	79.9	0.252	0.195	24.1	25.5	53.5	66.7
CVSE++ (Res152)	0.292	0.263	11.4	39.0	67.1	79.1	0.267	0.264	29.1	25.1	52.8	65.5
CVSE++ (Res152, Auto)	0.293	0.235	11.5	38.8	65.8	79.0	0.283	0.226	30.0	24.4	52.5	65.1
VSE++ (Res152, FT)	0.253	0.099	8.6	41.4	72.0	82.5	0.252	0.101	24.9	31.4	60.7	73.5
CVSE++ (Res152, FT, [10])	0.294	0.203	7.6	43.3	74.3	83.9	0.283	0.217	16.9	32.2	62.0	74.5
CVSE++ (Res152, FT)	0.300	0.301	8.3	43.0	73.6	83.5	0.288	0.290	20.7	31.4	61.1	73.8
CVSE++ (Res152, FT, Auto)	0.302	0.289	8.1	42.6	72.5	83.2	0.288	0.287	21.1	31.3	60.8	73.3

IV. EXPERIMENTS

A. Datasets & Implementation Detail

Following related works, Flickr30K [49] and MS-COCO [50], [51] 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 [5]. For MS-COCO, we also follow [5] 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 [5]. Our experimental settings follow that in VSE++ [5], which is the state-of-the-art for visual-semantic embedding. Note, in terms of image-sentence matching, SCAN [52] 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 [53] or ResNet-152 [54]-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 [5]. 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.

TABLE II
PERFORMANCE OF DIFFERENT METHODS ON STS-B BENCHMARK.

Method	Score on STS-B
CBoW	0.586
BERT [40]	0.865
BERT (Reproduced) [10]	0.880
Hybrid (CBoW+BERT) [10]	0.790
Sentence-BERT [11]	0.861

When evaluating the CS@K scores, we exclusively use the relevance degrees to determine the reference ranking list. 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 [55]. When being manually selected, the ladder number L is set as 2 if not mentioned, while the threshold θ_1 for splitting $\mathcal{N}_1^{-q}$ and $\mathcal{N}_2^{-q}$ is fixed at 0.63 [10] for the "hybrid" relevance degree, or at 0.40 for the Sentence-BERT relevance degree. Accordingly, the margins and the loss weights are set as $\alpha_1 = 0.2$, $\alpha_2 = 0.01$, $\beta_1 = 1$, $\beta_2 = 0.25$, respectively. When the ladder number L is selected adaptively, its lower and upper bound (L_{MIN}, L_{MAX}) are set as $(2, 4)$ otherwise mentioned. Accordingly, the margins are set as $\alpha_1 = 0.2$ and $\alpha_i = 0.01 (i \in \{2, 3, 4\})$, while the loss weights are set as

$\beta_1 = 1.0$ and $\beta_i = 1/2^i (i \in \{2, 3, 4\})$.

B. Relevance Degree

As pointed out by [11], [10], the BERT inference is highly computational expensive. To alleviate this, we introduce a "hybrid" mechanism by combining CBoW and BERT in the conference version of this paper [10]. Although it achieves a relatively high performance on the standard benchmark, the fact that the relevance degrees by CBoW and that of BERT follow different distributions means such "hybrid" supervision information may introduce noise into the training process. In this paper, we employ Sentence-BERT [11] as described in Section III-A. It greatly reduces the effort for finding the similarity matrices for our sentences, while maintaining the accuracy from a fine-tuned BERT.

Specifically, the performance of these methods on the STS-B benchmark is summarized in Table. II. In terms of learning a coherent visual semantic embedding space, Sentence-BERT is expected to be effective due to its competitive performance on this benchmark. Moreover, it provides more accurate relevance degrees compared to the "hybrid" mechanism used in [10].

Time consumption for calculating relevance degrees differs across these methods. Since the mini-batch size is set to 128 throughout all experiments, every iteration in the training process of CVSE++ involves the calculation of a pairwise relevance degree matrix of size 128×128. We have measured the time consumption of every method used in our experiments for such relevance degree matrix calculation with Python cProfile. Our hardware platform consists of two Intel Xeon 6226R CPUs and 8 Nvidia RTX3090 GPUs. During our experiments, the average time consumption for an 128×128 relevance degree is 7.38×10^{-4} second with CBoW; 1.01×10^{-3} second with Sentence-BERT (the sentence representation vectors are pre-calculated); 1.18×10^{1} with "hybrid" CBoW+BERT mechanism [10]. As discussed in [10], solely using BERT for relevance degree is infeasible due to excessive time consumption. In contrast, the VSE++ method does not need such matrix, and hence can finish the training epochs in relatively less time. However, CVSE++ achieves a much better coherence in the visual semantic embedding space at an acceptable computation cost.

C. 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++ [5] 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 [5].

In our approach, the ladder number L can be decided either manually or adaptively. When manually specifying the ladder number L in the loss function, it 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.*, $L^2_{lad}(q)$ 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.*, $L^3_{lad}(q)$ 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 0.40, 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 Table I, where the proposed CVSE++ approaches evidently outperform their corresponding VSE++ counterparts in terms of CS@K, *e.g.*, from VSE++(Res152): 0.264 to CVSE++(Res152): 0.301 in terms of CS@100 for image$\rightarrow$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 3-fold increase over "VSE++ (Res152,FT)" (from 0.099 to 0.303) in image$\rightarrow$sentence retrieval. We also provide results of models trained using the relevance degree in the conference version [10] but evaluated using the relevance degree discussed in Section III-A, *e.g.*, that of "CVSE++ (Res152, [10])". As expected, since the adaptive ladder loss can better exploit the training data, it achieves higher CS@100 in most cases, for instance, from 0.316 of CVSE++ (VGG19,FT) to 0.331 of CVSE++ (VGG19,FT,Auto) for image-to-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. Moreover, the proposed adaptive ladder loss could further boost the coherence of the learned embedding space in a local scope (*i.e.*, in terms of CS@100), while the ladder loss with a manually selected ladder level L is still better at maintaining coherence from the global scope (*i.e.* in terms of CS@1000).

Simultaneously, a less expected phenomenon can be ob-

TABLE III
COMPARISON BETWEEN VSE++ AND CVSE++ IN TERMS OF CS@K AND R@K ON FLICKR30K DATASET.

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)	0.160	0.172	17.4	40.6	68.7	78.3	0.153	0.157	28.3	28.1	58.0	69.9
CVSE++ (VGG19, [10])	0.162	0.158	16.5	42.5	69.5	79.3	0.160	0.171	25.8	28.4	58.8	70.6
CVSE++ (VGG19)	0.167	0.215	19.2	40.2	68.9	78.2	0.161	0.206	28.6	27.6	57.9	68.8
CVSE++ (VGG19, Auto)	0.170	0.184	19.7	39.7	67.3	77.5	0.161	0.175	29.9	27.3	57.5	69.4
VSE++ (VGG19, FT)	0.169	0.177	13.9	44.7	73.2	81.7	0.166	0.176	22.2	32.1	63.2	74.1
CVSE++ (VGG19, FT, [10])	0.178	0.183	12.5	44.8	73.8	84.3	0.173	0.183	19.8	34.6	65.7	76.5
CVSE++ (VGG19, FT)	0.187	0.222	14.1	43.8	72.6	81.0	0.173	0.219	22.5	32.9	63.3	73.4
CVSE++ (VGG19, FT, Auto)	0.189	0.211	15.2	43.1	72.1	82.3	0.178	0.205	21.8	32.3	64.0	74.6
VSE++ (Res152)	0.165	0.166	10.8	48.9	77.8	86.5	0.160	0.165	20.6	36.0	65.8	75.4
CVSE++ (Res152, [10])	0.166	0.178	9.1	50.9	79.6	87.8	0.161	0.178	19.8	37.3	67.1	76.7
CVSE++ (Res152)	0.168	0.216	11.3	48.9	77.8	86.7	0.163	0.210	21.0	36.1	66.3	75.5
CVSE++ (Res152, Auto)	0.171	0.195	10.7	49.3	77.2	86.3	0.164	0.188	21.8	36.0	65.4	76.2
VSE++ (Res152, FT)	0.173	0.189	7.8	54.4	81.2	89.3	0.166	0.190	16.5	40.2	70.1	79.6
CVSE++ (Res152, FT, [10])	0.175	0.194	6.7	56.2	82.4	90.8	0.167	0.196	15.4	42.7	71.8	80.3
CVSE++ (Res152, FT)	0.183	0.211	7.7	56.5	82.8	90.1	0.172	0.214	15.9	41.7	71.1	80.0
CVSE++ (Res152, FT, Auto)	0.186	0.235	8.1	57.7	84.7	91.1	0.176	0.234	15.5	41.9	72.7	81.5

TABLE IV
PERFORMANCE OF THE PROPOSED CVSE++(RES152) WITH RESPECT TO THE PARAMETER β_2 (ON MS-COCO DATASET).

Ladder Selection	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
Manual	0.00	0.264	0.107	3.0	63.4	87.7	95.6	0.263	0.112	7.7	47.6	80.3	90.1
Manual	0.25	0.301	0.265	3.1	65.2	88.6	95.0	0.292	0.266	6.9	48.3	79.8	89.3
Manual	0.50	0.328	0.269	4.0	55.5	84.0	91.7	0.283	0.287	8.1	42.6	76.7	87.7
Auto (2, 2)	0.00	0.264	0.107	3.0	63.4	87.7	95.6	0.263	0.112	7.7	47.6	80.3	90.1
Auto (2, 2)	0.25	0.302	0.239	3.2	60.3	87.3	95.1	0.296	0.230	7.2	45.4	79.4	89.7
Auto (2, 2)	0.50	0.300	0.205	6.5	49.4	75.7	84.6	0.300	0.223	9.0	36.9	72.0	83.7

TABLE V
PERFORMANCE OF THE PROPOSED CVSE++(RES152) WITH RESPECT TO THE PARAMETER β_2 (ON FLICKR30K DATASET).

Ladder Selection	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
Manual	0.00	0.165	0.166	10.8	48.9	77.8	86.5	0.160	0.165	20.6	36.0	65.8	75.4
Manual	0.25	0.168	0.216	11.3	48.9	77.8	86.7	0.163	0.210	21.0	36.1	66.3	75.5
Manual	0.50	0.185	0.211	15.4	43.4	72.2	81.8	0.162	0.219	21.9	33.6	64.1	74.8
Auto (2, 2)	0.00	0.165	0.166	10.8	48.9	77.8	86.5	0.160	0.165	20.6	36.0	65.8	75.4
Auto (2, 2)	0.25	0.170	0.192	11.3	49.2	75.7	85.9	0.166	0.187	21.3	35.2	65.4	75.9
Auto (2, 2)	0.50	0.163	0.198	14.6	37.3	67.8	79.6	0.167	0.172	29.5	26.9	58.0	70.0

served from Table I: 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): 63.4 to CVSE++(Res152): 65.2 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.

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 Figure 4 (sentence→image). Similarly, we also randomly sample several images from the dataset, and showcase the top-30 retrieved sentences in Figure 5 (image→sentence). From the figures, it is clear that both methods, *i.e.*, VSE++ and CVSE++ (ours) can rank the ground-truth (totally-relevant) sample on the very top part of the list. However, VSE++ treats non-ground-truth samples as totally-irrelevant candidates and learns an incoherent embedding space, hence resulting in obviously more totally-irrelevant samples in the top-30 retrieval list. In contrast, our CVSE++ retrieves less totally-irrelevant samples, and most of the retrieved samples are still somewhat related to the ground-truth. For instance, in the first example of Figure 4 where the sentence is "a herd of elephants standing on the side of a grass covered hill", the retrieval results of VSE++ includes images of food (*e.g.*, donuts), snow-covered mountains, trains, *etc.*, which are totally-irrelevant to the query sentence. With the same query sentence, nearly all the retrieval results from CVSE++ match at least one of the following keys: "elephants", "animal standing somewhere", and "grass hill". Unlike VSE++, our learned embedding space is more coherent.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3139210, IEEE Transactions on Multimedia

TABLE VI
PERFORMANCE OF THE PROPOSED CVSE++(RES152) WITH RESPECT TO THE LADDER NUMBER L. (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	0.264	0.214	0.107	3.0	63.4	87.7	95.6	0.263	0.216	0.112	7.7	47.6	80.3	90.1
2	0.301	0.273	0.265	3.1	65.2	88.6	95.0	0.292	0.251	0.266	6.9	48.3	79.8	89.3
3	0.298	0.288	0.278	3.7	62.0	88.0	94.4	0.279	0.260	0.281	7.7	46.8	78.5	89.0

TABLE VII
PERFORMANCE OF THE PROPOSED CVSE++(RES152) WITH RESPECT TO THE LADDER NUMBER L. (ON FLICKR30K 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	0.165	0.151	0.166	10.8	48.9	77.8	86.5	0.160	0.149	0.165	20.6	36.0	65.8	75.4
2	0.168	0.161	0.216	11.3	48.9	77.8	86.7	0.163	0.156	0.210	21.0	36.1	66.3	75.5
3	0.172	0.175	0.216	12.5	47.7	76.4	85.3	0.160	0.164	0.215	21.1	35.7	65.6	75.4

Similarly, the first example in Figure 5 also suggest that nearly all the retrieval results from our coherent embeddings match at least one of these key words: "elephant", "animal standing somewhere", "grass hill or grass field". Evidently, our CSVE++ can put more somewhat-relevant candidates and reduce the number of totally-irrelevant candidates on the top-N retrieval list and enhance user experience.

D. Results on Flickr30K

Our approach is also evaluated on the Flikr30K dataset and compared with the baseline VSE++ variants, as shown in Table III. The hyper-parameter settings are identical to that in Table I with MS-COCO (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.

V. PARAMETER SENSITIVITY ANALYSIS AND DISCUSSIONS

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), the ladder number L, as well as the $(L_{\text{MIN}}, L_{\text{MAX}})$ parameters.

A. 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 Table IV and Table V, we can see that CS@K metrics improve with larger β_2, but R@K metrics degrade when β_2 is close to 0.5. Based on the three β_2 settings in Table IV and Table V, we speculate that CS@K and R@K metrics would not necessarily peak simultaneously at the same β_2 value for both manually selected L (denoted as "Manual") and adaptively selected one (denoted as "Auto (2,2)"). 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.

B. 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 $L^2_{lad}(q)$ is responsible for the optimization of the ranking order of very top candidates, *e.g.*, from top-1 $\sim$ top-100; while the high-level ladder $L^3_{lad}(q)$ is responsible for the optimization of the ranking order of subsequent candidates, *e.g.*, from top-100 $\sim$ top-200. Inevitably, a very large ladder number results in high 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 θ_2, α_3, β_3 are empirically fixed at 0.35, 0.01, 0.125, respectively. The experimental results are summarized in Table VI and Table VII. 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3139210, IEEE Transactions on Multimedia

TABLE VIII
PERFORMANCE OF THE PROPOSED CVSE++(RES152, AUTO) WITH RESPECT TO THE K-MEANS LOWER/UPPER BOUNDS. (ON MS-COCO DATASET)

(L_{MIN}, L_{MAX})	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
(1, 1)	0.264	0.107	3.0	63.4	87.7	95.6	0.263	0.112	7.7	47.6	80.3	90.1
(2, 2)	0.302	0.239	3.2	60.3	87.3	95.1	0.296	0.230	7.2	45.4	79.4	89.7
(3, 3)	0.301	0.253	3.4	60.1	87.8	94.1	0.295	0.247	7.5	44.2	78.6	88.9
(4, 4)	0.306	0.253	3.1	60.8	88.0	94.4	0.297	0.250	7.4	45.2	79.1	88.8
(5, 5)	0.301	0.253	3.3	60.4	88.5	94.9	0.298	0.246	7.5	45.6	79.2	89.3
(2, 3)	0.303	0.238	3.3	61.8	87.2	95.0	0.295	0.229	7.3	46.4	79.2	89.4
(2, 4)	0.305	0.236	3.1	64.7	88.1	95.2	0.295	0.229	7.1	47.2	79.1	89.5
(2, 5)	0.300	0.237	3.4	59.6	88.7	94.6	0.295	0.229	7.2	46.8	78.9	89.3

TABLE IX
PERFORMANCE OF THE PROPOSED CVSE++(RES152, AUTO) WITH RESPECT TO THE K-MEANS LOWER/UPPER BOUNDS. (ON FLICKR30K DATASET)

(L_{MIN}, L_{MAX})	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
(1, 1)	0.165	0.166	10.8	48.9	77.8	86.5	0.160	0.165	20.6	36.0	65.8	75.4
(2, 2)	0.170	0.192	11.3	49.2	75.7	85.9	0.166	0.187	21.3	35.2	65.4	75.9
(3, 3)	0.162	0.203	11.5	48.1	74.9	84.2	0.165	0.196	21.8	33.0	64.2	74.8
(4, 4)	0.166	0.205	10.9	47.7	76.0	85.6	0.167	0.199	21.5	35.0	65.4	75.6
(5, 5)	0.168	0.201	10.8	47.9	75.2	85.4	0.166	0.194	21.6	36.1	65.1	75.9
(2, 3)	0.168	0.196	10.4	49.1	76.1	85.8	0.166	0.190	20.3	35.6	65.9	76.4
(2, 4)	0.171	0.195	10.7	49.3	77.2	86.3	0.164	0.188	21.8	36.0	65.4	76.2
(2, 5)	0.164	0.197	12.2	48.2	76.7	84.9	0.163	0.193	20.1	35.4	65.3	76.0

TABLE X
ADAPTIVE LADDER LOSS WITH HYBRID (CBOW+BERT) RELEVANCE DEGREE ON MS-COCO DATASET (1000 TEST SAMPLES).

Model	MS-COCO Dataset (1000 Test Samples)											
	Image to Sentence						Sentence to 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
CVSE++ (VGG19, [10])	0.281	0.175	3.7	56.5	85.1	92.9	0.273	0.178	7.3	43.0	78.8	88.7
CVSE++ (VGG19, [10], Auto)	0.295	0.145	4.0	56.2	84.5	92.6	0.294	0.148	8.8	42.9	78.0	88.0
CVSE++ (Res152, [10])	0.290	0.192	2.9	66.5	89.0	95.2	0.278	0.196	6.0	48.6	80.5	90.5
CVSE++ (Res152, [10], Auto)	0.302	0.166	3.1	64.3	88.3	94.4	0.292	0.170	7.2	46.9	79.2	89.6
Flickr30K Dataset												
CVSE++ (VGG19, [10])	0.162	0.158	16.5	42.5	69.5	79.3	0.160	0.171	25.8	28.4	58.8	70.6
CVSE++ (VGG19, [10], Auto)	0.166	0.144	16.7	39.5	67.6	78.4	0.161	0.146	29.4	27.8	57.1	68.6
CVSE++ (Res152, [10])	0.166	0.178	9.1	50.9	79.6	87.8	0.161	0.178	19.8	37.3	67.1	76.7
CVSE++ (Res152, [10], Auto)	0.169	0.151	11.1	48.8	77.6	85.8	0.167	0.166	23.0	35.6	64.9	74.5

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.

C. Adaptive Ladder Loss

Manually selected ladder level (and fixed thresholds) are not flexible enough to better exploit the training data, while adaptive ladder loss can further enhance the coherence of the top part (local scope) of the ranking list, as the β_i weight for the first several ladders are relatively larger.

In this part, we study the selection of the bounds of K-Means clustering for adaptive ladder loss, *i.e.*, $(L_{\text{MIN}}, L_{\text{MAX}})$. The experimental results can be found in Table. VIII and Table. IX. As shown in the upper parts of the tables, when $L_{\text{MIN}} = L_{\text{MAX}}$, the ladder level L is fixed, while the thresholds separating the samples are adaptively determined. In this case the model achieves clearly better coherence compared to the VSE++ baseline (*i.e.*, $L_{\text{MIN}} = L_{\text{MAX}} = 1$, which means the adaptively determined inequality chain is effective in improving the embedding coherence. We also note that the model performance is not sensitive to the exact number of ladders. In the lower parts of Table. VIII and Table. IX, we endow ladder loss with more flexibility to select a proper L, and observe that the model performs the best with $(2, 4)$, which is better than the model with the $(2, 2)$ setting. This means the flexibility to determine a proper L is beneficial for the model. Compared to using manually a selected L and thresholds, adaptive ladder loss is better at coherence in a local scope.

D. Adaptive Ladder Loss with Hybrid Relevance Degree

To better demonstrate the effectiveness of the proposed adaptive ladder loss, we also perform experiments with the "hybrid" relevance degree used in the conference version of this paper [10], as shown in Table. X. For instance, comparing the results of "CVSE++ (Res152, [10])" and "CVSE++

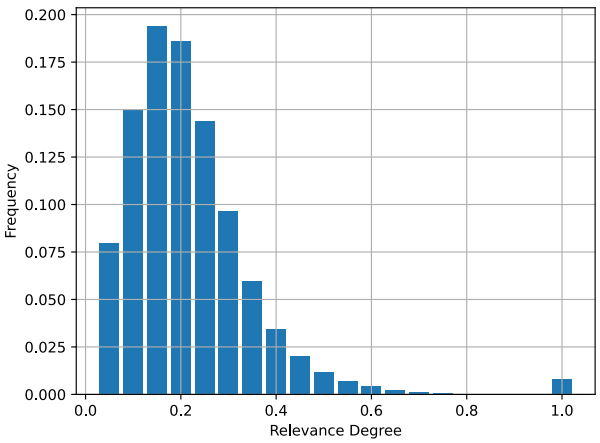


Fig. 6. Histogram of Elements in Relevance Degree Matrices during the Training Process on MS-COCO Dataset.

TABLE XI
COMPARISON OF MAP PERFORMANCE ON MS-COCO (1000 TEST SAMPLES) AND FLICKR30K DATASETS.

Model	mAP (Sentence to Image)	
	MS-COCO	Flickr30K
⌟VSE++ (VGG19)	0.428	0.265
CVSE++ (VGG19, [10])	0.433	0.271
CVSE++ (VGG19)	0.440	0.277
CVSE++ (VGG19, Auto)	0.443	0.282
⌟VSE++ (VGG19, FT)	0.435	0.308
CVSE++ (VGG19, FT, [10])	0.486	0.310
CVSE++ (VGG19, FT)	0.489	0.317
CVSE++ (VGG19, FT, Auto)	0.495	0.325
⌟VSE++ (Res152)	0.463	0.342
CVSE++ (Res152, [10])	0.471	0.342
CVSE++ (Res152)	0.472	0.343
CVSE++ (Res152, Auto)	0.480	0.349
⌟VSE++ (Res152, FT)	0.522	0.381
CVSE++ (Res152, FT, [10])	0.525	0.382
CVSE++ (Res152, FT)	0.525	0.389
CVSE++ (Res152, FT, Auto)	0.540	0.395

(Res152, [10], Auto)", a higher CS@100 is achieved by the adaptive ladder loss. Thus, the proposed adaptive ladder loss is still effective for improving the visual-semantic embedding coherence with a different relevance degree.

E. Histogram of Relevance Degrees

As aforementioned, the ladder loss with fixed thresholds [10] can be inefficient in learning in a coherent embedding space. As shown in Figure. 6, we create a histogram for the elements in relevance degree matrices (using Sentence-BERT) during the first 500 epochs during the training process on the MS-COCO dataset. From the figure, we note that there is a large variance in the relevance degrees for a batch of samples. This means the proposed adaptive ladder loss is expected to be more effective in learning a coherent embedding space, compared to that with fixed thresholds which are not flexible enough with random mini-batches.

F. mAP Performance for Image-to-Sentence Retrieval

The mean average precision (mAP) metric is also able to partly reflect the coherence of the embedding space. In the datasets adopted in our experiments, each image is assigned with 5 corresponding descriptive sentences. In this case, a well-trained model is expected to rank all the corresponding sentences ahead of as many other examples as possible, and hence leading to a higher mAP performance. Although mAP can partly reflect the coherence of embedding space, it is still limited because it also treats the samples in a bi-polar way. Further more, this performance metric is only applicable for the image-to-sentence retrieval task.

Following these, we measure the mAP performance of models in Table. I and Table III, and summarize it in Table XI. According to the results, the ladder loss function is beneficial for improving the mAP performance for image-to-sentence retrieval (*e.g.*, by comparing the "VSE++ (Res152)" model and the "CVSE++ (Res152, [10])" model). With a better relevance degree (*i.e.*, Sentence-BERT), the mAP performance can be improved. Last but not least, since the model more efficiently leverages the relevance degree information with adaptive ladder loss (*e.g.*, by comparing "CVSE++ (Res152)" and "CVSE++ (Res152, Auto)" model), the mAP performance can be further boosted. A higher mAP performance means the model is able to consistently map sentences corresponding to the same images closer to the image, which is also expected in a coherent embedding space. Thus, the improvement in mAP also reflects that our proposed adaptive ladder loss further enhances the coherence of embedding space.

VI. 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. Moreover, the inequality chain underlying the ladder loss can be determined in an adaptive way to further boost the coherence of the embedding space. 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.

REFERENCES

[1] X. Ji, W. Wang, M. Zhang, and Y. Yang, "Cross-domain image retrieval with attention modeling," in *ACM Int. Conf. Multimedia (ACM MM)*, 2017, pp. 1654–1662.
[2] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *Proc. Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2014.
[3] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2015, pp. 3128–3137.

[4] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image-sentence mapping," in *Proc. Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2014, pp. 1889–1897.

[5] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2015, pp. 815–823.

[7] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2016, pp. 1857–1865.

[8] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2006, pp. 1735–1742.

[9] D. Downey, S. Dumais, and E. Horvitz, "Heads and tails: studies of web search with common and rare queries," in *ACM Spec. Inter. Group on Info. Retriev. (SIGIR)*, 2007, pp. 847–848.

[10] M. Zhou, Z. Niu, L. Wang, Z. Gao, Q. Zhang, and G. Hua, "Ladder loss for coherent visual-semantic embedding," in *Proc. AAAI. Conf. Artif. Intell. (AAAI)*, 2020, pp. 13 050–13 057.

[11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese bert-networks," in *Proc. Conf. on Empir. Meth. in Natural Lang. Proc. (EMNLP)*, 2019, pp. 3982–3992.

[12] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1–9.

[13] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[14] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. J. Comput. Vis. (IJCV)*, vol. 106, no. 2, pp. 210–233, 2014.

[15] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 529–545.

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

[17] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 1247–1255.

[18] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2015, pp. 3441–3450.

[19] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Ranzato, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2013, pp. 2121–2129.

[20] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 5005–5013.

[21] R. Socher, Q. Le, C. Manning, and A. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Trans. of the Assoc. for Comp. Ling. (TACL)*, vol. 2, pp. 207–218, 2014.

[22] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Hierarchical multimodal lstm for dense visual-semantic embedding," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1899–1907.

[23] Q. Lin, W. Cao, Z. He, and Z. He, "Mask cross-modal hashing networks," *IEEE Trans. Multimedia*, vol. 23, pp. 550–558, 2021.

[24] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3196–3209, 2020.

[25] X. Ma, T. Zhang, and C. Xu, "Multi-level correlation adversarial hashing for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3101–3114, 2020.

[26] C. Zhao, X. Lv, Z. Zhang, W. Zuo, J. Wu, and D. Miao, "Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3180–3195, 2020.

[27] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 4004–4012.

[28] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krähenbühl, "Sampling matters in deep embedding learning," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2859–2867.

[29] Y. Yuan, K. Yang, and C. Zhang, "Hard-aware deeply cascaded embedding," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 814–823.

[30] K. Roth, T. Milbich, S. Sinha, P. Gupta, B. Ommer, and J. P. Cohen, "Revisiting training strategies and generalization performance in deep metric learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 8212–8222.

[31] C. Huang, C. C. Loy, and X. Tang, "Local similarity-aware deep feature embedding," in *Proc. Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2016, pp. 1262–1270.

[32] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," in *Proc. Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2016, pp. 4170–4178.

[33] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 5022–5030.

[34] M. Zhou, Z. Niu, L. Wang, Q. Zhang, and G. Hua, "Adversarial ranking attack and defense," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 781–799.

[35] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked list loss for deep metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 5202–5211.

[36] H. Xuan, A. Stylianou, X. Liu, and R. Pless, "Hard negative examples are hard, but useful," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 126–142.

[37] H. Jun, B. Ko, Y. Kim, I. Kim, and J. Kim, "Combination of multiple global descriptors for image retrieval," *arXiv preprint arXiv:1903.10663*, 2019.

[38] W. Min, S. Mei, Z. Li, and S. Jiang, "A two-stage triplet network training framework for image retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3128–3138, 2020.

[39] B. Ionescu, M. Rohm, B. Boteanu, A. L. Gînscă, M. Lupu, and H. Müller, "Benchmarking image retrieval diversification techniques for social media," *IEEE Trans. Multimedia*, vol. 23, pp. 677–691, 2021.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Americ. Chap. Assoc. Comput. Ling. (NAACL)*, Jun. 2019, pp. 4171–4186.

[41] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. North Americ. Chap. Assoc. Comput. Ling. (NAACL)*, 2018, pp. 2227–2237.

[42] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *Proc. Assoc. Comp. Ling. (ACL)*, Jul. 2019, pp. 4487–4496.

[43] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *Proc. Conf. on Empir. Meth. in Natural Lang. Worksh. (EMNLPW)*, pp. 353–355, 2018.

[44] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation," in *Proc. Int. Worksh. on Sem. Eval.*, 2017, pp. 1–14.

[45] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li, "On the sentence embeddings from pre-trained language models," in *Proc. Conf. on Empir. Meth. in Natural Lang. Proc. (EMNLP)*, 2020, pp. 9119–9130.

[46] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 118–126.

[47] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[48] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

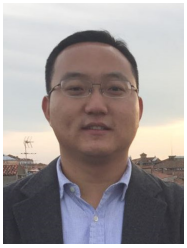
[49] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2641–2649.

[50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

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

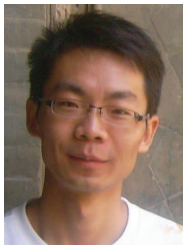
[52] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 201–216.

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 770–778.
[55] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. Adv. Neural Inform. Process. Syst. Worksh. (NeurIPSW)*, 2017.

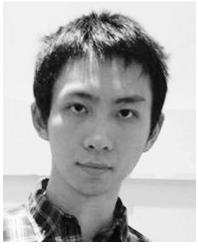


Le Wang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in Control Science and Engineering from Xi'an Jiaotong University, Xi'an, China, in 2008 and 2014, respectively. From 2013 to 2014, he was a visiting Ph.D. student with Stevens Institute of Technology, Hoboken, New Jersey, USA. From 2016 to 2017, he is a visiting scholar with Northwestern University, Evanston, Illinois, USA. He is currently an Associate Professor with the Institute of Artificial Intelligence and Robotics of Xi'an Jiaotong University, Xi'an, China. His research interests include computer vision, pattern recognition, and machine learning. He is the author of more than 60 peer reviewed publications in prestigious international journals and conferences. He is an area chair of CVPR'2022.

Mo Zhou (Student Member, IEEE) received the B.S. degree in Electromagnetic Fields and Wireless Technology, and the M.S. degree in Pattern Recognition and Intelligent System from Xidian University, Xi'an, China, in 2017 and 2020. He is currently a research assistant with the Institute of Artificial Intelligence and Robotics of Xi'an Jiaotong University, Xi'an, China. His research interests include deep learning, computer vision, and adversarial attack and defense.



Zhenxing Niu (Member, IEEE) received the Ph.D. degree in Control Science and Engineering from Xidian University, Xi'an, China, in 2012. From 2013 to 2014, he was a visiting scholar with University of Texas at San Antonio, Texas, USA. He is a Researcher at Alibaba Group, Hangzhou, China. Before joining Alibaba Group, he is an Associate Professor of School of Electronic Engineering at Xidian University, Xi'an, China. His research interests include computer vision, machine learning, and their application in object discovery and localization. He served as PC member of CVPR, ICCV, and ACM Multimedia.



Qilin Zhang (Member, IEEE) received the B.E. degree in Electrical Information Engineering from the University of Science and Technology of China, Hefei, China, in 2009, and the M.S. degree in Electrical and Computer Engineering from University of Florida, Gainesville, Florida, USA, in 2011, and the Ph.D. degree in Computer Science from Stevens Institute of Technology, Hoboken, New Jersey, USA, in 2016. He is currently a Lead Research Engineer with Here Technologies, Chicago, Illinois, USA. His research interests include computer vision, machine learning and multimedia signal processing. He is the author of more than 40 peer reviewed publications in prestigious international journals and conferences.

Nanning Zheng (Fellow, IEEE) graduated from the Department of Electrical Engineering of Xi'an Jiaotong University (XJTU), Xi'an, China, in 1975, received the M.E. degree in Information and Control Engineering from Xi'an Jiaotong University, Xi'an, China, in 1981, and the Ph. D. degree in electrical engineering from Keio University, Keio, Japan, in 1985. He is currently a Professor and the Director with the Institute of Artificial Intelligence and Robotics of Xi'an Jiaotong University, Xi'an, China. His research interests include computer vision, pattern recognition, computational intelligence, and hardware implementation of intelligent systems. Since 2000, he has been the Chinese representative on the Governing Board of the International Association for Pattern Recognition. He became a member of the Chinese Academy Engineering in 1999.