Uncertainty-based Cross-Modal Retrieval with Probabilistic Representations

Leila Pishdad[1] Ran Zhang[1] Konstantinos G. Derpanis[1,2] Allan Jepson[1,3]
Afsaneh Fazly[1]
[1]Samsung AI Centre Toronto, [2]York University, [3]University of Toronto
leila.pishdad@mail.mcgill.ca , kosta@yorku.ca, {ran.zhang, allan.jepson, a.fazly}@samsung.com

Abstract

Probabilistic embeddings have proven useful for capturing polysemous word meanings, as well as ambiguity in image matching. In this paper, we study the advantages of probabilistic embeddings in a cross-modal setting (i.e., text and images), and propose a simple approach that replaces the standard vector point embeddings in extant image–text matching models with probabilistic distributions that are parametrically learned. Our guiding hypothesis is that the uncertainty encoded in the probabilistic embeddings captures the cross-modal ambiguity in the input instances, and that it is through capturing this uncertainty that the probabilistic models can perform better at downstream tasks, such as image-to-text or text-to-image retrieval. Through extensive experiments on standard and new benchmarks, we show a consistent advantage for probabilistic representations in cross-modal retrieval, and validate the ability of our embeddings to capture uncertainty.

1. Introduction

In this paper, we consider the challenge of cross-modal retrieval with focus on relating image and text modalities. The general aim is to learn a shared representational space, where related instances from multiple modalities are mapped closely together in the joint space. Given the mappings and a query element from one domain, retrieval amounts to finding the nearest (i.e, most relevant) elements in a database in the other domain.

Most cross-modal retrieval methods (e.g., [10, 23]) learn mappings of each data item (e.g., an image–caption pair) to a vector representing a point in a k-dimensional space. A major issue with these point embeddings is their expressiveness in terms of the information they can capture about the inputs. For language, a clear drawback of point representations is their inability to handle polysemy, i.e., words or phrases having multiple meanings [41]. For images, uncertainty can be introduced by the image formation process (e.g., occlusion) [19]. For image–text matching, an image can be described in many ways, and a caption can describe many different images, resulting in some degree of cross-modal ambiguity. Point embeddings lack the ability to represent the uncertainties in cross-modal associations and skew to learning an average of the cross-modal information by regressing to the mean embedding.

In this work, we forgo point mappings and instead relate inputs to a region of the shared representation space. The inputs from each modality are modeled as probability distributions (i.e., probabilistic embeddings) in a common embedding space. In particular, the inputs are mapped to Gaussian distributions, where the distribution spread in the joint embedding space captures the mapping's uncertainty (or ambiguity). While few works have explored similar embeddings in a single modality (i.e., text [41] or images [32]) their extension in the cross-modal setting is underexplored with only a single concurrent study [9] conducted to date.

Contributions. We introduce a simple and general approach to learning probabilistic embeddings (instead of point embeddings) in an image–text matching model, that can explicitly handle and express the inherent uncertainty in establishing cross-modal associations. Through a comprehensive empirical study of several extant retrieval models and benchmarks, we show a consistent advantage of learning probabilistic embeddings for cross-modal retrieval over their point-based counterparts. Moreover, we show (via a controlled experiment) that a per-instance measure of uncertainty actually captures the cross-modal ambiguity, which we attribute as a key factor for the consistently high performance of the probabilistic models. We view our contribution as complementary to other innovations in the cross-modal retrieval space (e.g., embedding architectures). To avoid conflating retrieval performance increases due to the representation (i.e., point vs. probabilistic) with differences in the loss function and/or the backbone architecture, we use standard retrieval methods [10, 23] in our evaluation and simply replace their terminal point mappings with our probabilistic ones. Empirically, our retrieval results exceed

or are competitive to the adapted baseline point embedding networks, while also providing the added benefit of a diagnostic uncertainty measure. In comparison to the lone recent probabilistic method to cross-modal retrieval [9], we demonstrate consistent superior performance under all performance metrics.

2. Related work

One-to-one mappings. For image-based retrieval, most methods are based on (handcrafted or learned) point-based representations of the imagery [2, 16, 17, 25, 35]. Similarly, standard cross-modal retrieval methods rely on (independent) one-to-one mappings that take elements from their respective domains (e.g., images or text) as a whole and map them to a shared embedding space, where similarities can be established. A popular approach to learn these mappings is based on a ranking loss (e.g., [10, 11, 38, 40]), where related elements are encouraged to be closer than unrelated ones. Alternative models aid learning of the shared embedding space through encoding intra-modality similarities with additional auxiliary losses [48] or leveraging adversarial learning [7]. As described in Sec. 1, one-to-one mappings cannot handle ambiguities (e.g., polysemy) in the inputs. We learn probabilistic embeddings that explicitly encode uncertainty through mapping each input instance to a region instead of a point in a shared representation space.

Many-to-many mappings. To better handle the inherent ambiguities in multimodal inputs, recent methods [22, 23, 26, 27, 31, 39, 42–44, 47] first generate multiple embeddings that capture parts of the input in each domain and then use a module to selectively combine the various parts. For instance, the model of [39] takes images and sentences, and first generates a set of image regions and word embeddings. These part embeddings are then combined *independently* in each domain using multi-head self-attention to generate multiple diverse representations for matching. Alternatively, a plethora of recent methods integrate cross-attention operating *jointly* on image region and word embeddings [15, 22, 26, 27, 31, 42, 43, 46, 47], i.e., mixing information across domains. While these latter methods are the state of the art in cross-modal retrieval, they require excessive computation that renders them inapplicable for real-world deployment: for each query, cross-attention is applied to each element in the search database. Our focus is on joint embeddings computed independently in each domain to facilitate large-scale, efficient search.

Probabilistic embeddings. Density-based or probabilistic embeddings have been previously used to capture words and their associated uncertainties in meanings [3–5, 29, 41]. Our method extends previous work [41] on Gaussian word embeddings to the multimodal joint embedding setting of images and captions. From these probabilistic embeddings, we learn a common latent space that captures nuances of meaning with respect to image–caption correspondences. Probabilistic embeddings have also been applied to images alone [32]. Concurrent work [9] extends this uni-modal (image-only) model to the same multimodal retrieval setting considered in our paper. Specifically, [9] combines ideas from a recent cross-modal architecture [39] with the hedged instance embedding approach of [32]. We propose a simpler approach that does not require a specialized loss function to estimate the instance distributions, and show benefits of learning probabilistic embeddings *independently* of the backbone architecture. As such, we integrate our approach directly with two popular cross-modal architectures and show that both benefit from probabilistic embeddings. Our extensive experimental results point to a clear advantage of our probabilistic embeddings over point-based embeddings, as well as the method of [9], across all performance metrics.

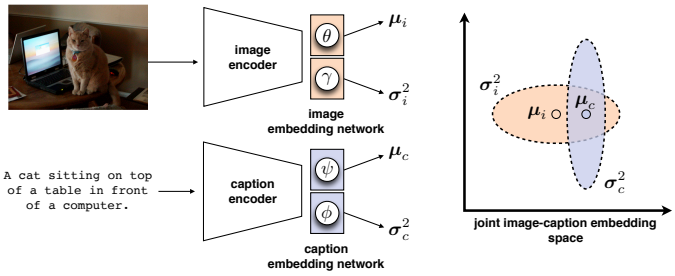


Figure 1. Representation overview. Each modality (image/caption) is mapped to a Gaussian distribution in the joint embedding space shown on the right.

3. Learning probabilistic representations

In this section, we describe our approach to learning density-based representations of images and captions in a joint embedding space. Fig. 1 provides an overview of our approach.

3.1. Model

We use Gaussian distributions, $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, with mean $\boldsymbol{\mu}$ and covariance Σ, to represent each image and caption in the joint embedding space. Following [41], we assume either a diagonal multivariate ellipsoidal or Gaussian distribution to model our probabilistic representations, and learn mappings to the mean and variance vectors. Let $f(\cdot)$ and $g(\cdot)$ denote the initial image and text encodings, respectively, and $f_{\text{emb}}(\cdot;\boldsymbol{\theta})$ and $g_{\text{emb}}(\cdot;\boldsymbol{\psi})$ represent the image and caption embedding functions, respectively, with learnable parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. We take the output of these functions to learn the means and use the same functions with different learnable parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ to learn the variances

of the Gaussian representations. Formally,

$$\boldsymbol{\mu}_i = f_{\text{emb}}(f(i); \boldsymbol{\theta}), \qquad \boldsymbol{\sigma}_i^2 = f_{\text{emb}}(f(i); \boldsymbol{\gamma}), \tag{1}$$

$$\boldsymbol{\mu}_c = g_{\text{emb}}(g(c); \boldsymbol{\psi}), \qquad \boldsymbol{\sigma}_c^2 = g_{\text{emb}}(g(c); \boldsymbol{\phi}), \tag{2}$$

where the means and variances $\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2, \boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2 \in \mathbb{R}^D$, and i and c denote the input image and caption, respectively.

With this formulation, our probabilistic approach can be easily adapted to point-based image–text models with minor modifications. Specifically, we can use the point-based embeddings from each modality as our mean outputs, and duplicate the learnable part of the architecture to learn the variances (see Fig. 1). We can then replace the vector-based similarity metric with a *probabilistic* one (see Section 3.2) without requiring any changes to the loss function, to learn the image and caption embeddings in the joint space. Please refer to Section 4.2 for a more detailed discussion.

3.2. Training loss

Our multimodal representation is trained using pairs of images and corresponding captions, (i_n, c_n) for $n \in \{1, \ldots, N\}$, where N is the total number of ground truth pairs in the dataset. A pair (i_j, c_k) is designated as *matching* if $j = k$ and *non-matching* if $j \neq k$. For ease of exposition, we use (i, c) to refer to the embeddings of a matching pair and (i', c) and (i, c') to represent the embeddings of a non-matching pair. Our model is trained using a contrastive loss to learn a joint (image and caption) embedding space in which the elements of a matching pair are closer to each other compared to those of non-matching pairs. For the contrastive loss, we use the hinge-based triplet ranking loss [6] with semi-hard negative mining [36]:

$$\begin{aligned}\mathcal{L}(i, c) = &\max_{c'} \left[\alpha + \text{sim}(i, c') - \text{sim}(i, c)\right]_+ \\ &+ \max_{i'} \left[\alpha + \text{sim}(i', c) - \text{sim}(i, c)\right]_+ ,\end{aligned} \tag{3}$$

where $[x]_+ \triangleq \max(\text{x}, 0)$, α is a hyperparameter for the margin, and $\text{sim}(\cdot, \cdot)$ measures the similarity between two distributions, i.e., the image and caption embeddings. The total loss is the sum of $\mathcal{L}(i, c)$ over all positive pairs in the batch, i.e., $\mathcal{L} = \sum_{(i,c)} \mathcal{L}(i, c)$.

There are a variety of similarity measures over distributions, $\text{sim}(\cdot, \cdot)$, that can be considered for our loss, (3). In this work, we consider three standard similarity measures, namely, negative Kullback-Liebler (KL) divergence, negative minimum KL divergence, and negative 2-Wasserstein distance; please see the supplemental for their definitions.

3.3. Per-instance measure of uncertainty

A key advantage of probabilistic representations is that they can express the uncertainty in the input. We propose an explicit measure of uncertainty for an input instance x (a caption or an image), calculated as the log determinant of the covariance matrix of the learned Gaussian distribution for x, i.e., $\log(\det(\Sigma_x))$. Based on this definition, the uncertainty of an instance corresponds to the overall variance of its learned distribution in the joint space. We hypothesize that this variance reflects the level of cross-modal ambiguity for an image/caption. That is, the lower the ambiguity of an instance, the lower the uncertainty of its learned representation in the joint space. We provide empirical evidence for this hypothesis (see Sec. 4.6), suggesting that it is indeed through an explicit capturing of the cross-modal ambiguity that probabilistic representations are able to perform more accurate cross-modal retrieval.

4. Empirical evaluation

4.1. Datasets

We report retrieval results on the standard MS-COCO image–caption dataset [24]. We use the splits from [18], containing $82,783$ training, 5000 validation, and 5000 test images, where each image is associated with five manually-provided captions. Following prior work [9,10,39], we augment our training data with $30,504$ images (and their captions) from the original validation set of MS-COCO that were left out in the split of [18]. We also report retrieval results on the smaller and noisier Flickr30K [45] dataset. Flickr30K contains $31,000$ images each paired with five manually-annotated captions. Following [22], we set aside 1000 image–caption pairs each for validation and test, and use the rest for training.

Prior work has noted a problem with MS-COCO for cross-modal retrieval evaluation: most pairs are considered as non-matching due to the sparseness of the annotations [13,14,33]. To address this issue, [33] introduced the Crisscrossed (CxC) dataset which extends MS-COCO with additional human annotations for semantic similarities on the validation and test sets. We use the CxC annotations on the test set as additional evaluation data for the models trained on MS-COCO. Following [33], we identify $10,614$ additional positive pairs based on the annotations, and add them to the original $25,000$ MS-COCO positive test pairs.

Finally, we evaluate on finer-grained retrieval using Visual Genome [21]. Visual Genome is a large, densely annotated image–caption dataset, with images containing tight bounding boxes around objects and a corresponding short caption describing the bounding box region. The dense annotations provide rich information about images, enabling us to investigate the relationship between the complexity of an image/caption and its representation uncertainty.

4.2. Implementation details

Our models are implemented in PyTorch [34]. To show the generality of our probabilistic representation, we adapt two distinct retrieval models: VSE++ [10] and

VSRN [23]. Both models learn image–caption embeddings *without* cross-domain attention. VSE++ is representative of a class of end-to-end models that learn a single embedding from the raw images/captions, and is widely used as a baseline for comparison in cross-modal retrieval. VSRN extracts a set of image features from regions identified by an object detector, and is among the best performing retrieval models that does not involve cross-domain attention. For both models, we use the official codebase.

As mentioned earlier, a key advantage of our probabilistic approach is that it can be adapted to any image–text model with simple modification to the original model. Specifically, we take the output of the original model as the mean. We learn the variance by duplicating the learnable part of the model and thus generate another output of the same dimension as the variance vector. Next, we describe our modifications to VSE++ and VSRN.

Image representations. For VSE++, to compute the mean, we linearly project the image features, $f(i) \in \mathbb{R}^{2048}$, generated by ResNet-152 [12], yielding $\mu_i \in \mathbb{R}^{1024}$. For VSRN, we use a recurrent network to consolidate the contextualized regions features, $f(i) \in \mathbb{R}^{36\times2048}$, generated by bottom-up attention [1] followed with a graph neural network. We take the final hidden state of the recurrent network as the mean vector, $\mu_i \in \mathbb{R}^{2048}$. As mentioned earlier, we duplicate the learnable embedding layers of the mean network with no parameter sharing to generate the variance vectors, yielding $\sigma_i \in \mathbb{R}^{1024}$ for VSE++ and $\sigma_i \in \mathbb{R}^{2048}$ for VSRN. Unless otherwise stated, we do not fine-tune the image backbones.

Caption representations. For our probabilistic versions of both VSE++ and VSRN, we use randomly initialized 300-dimensional embeddings to represent each word in a caption. Word embeddings are then passed through a unidirectional Gated Recurrent Unit (GRU) [8]. The final hidden state of the GRU is taken as the caption mean embedding in the joint space, with $\mu_c \in \mathbb{R}^{1024}$ for VSE++ and $\mu_c \in \mathbb{R}^{2048}$ for VSRN. For the variances, we duplicate the GRU layer with no parameter sharing in both VSE++ and VSRN.

Training details. We train all our models for 30 epochs with a learning rate of $2\mathrm{e}{-4}$ and a learning rate decay factor of 10 at epoch 15. We use the Adam optimizer [20] with a mini-batch size of 128. For the loss function, (3), we use a margin of 0.2 for VSE++ and 0.1 for VSRN. To ensure numerical stability, we estimate the log variances instead of variances, and following previous work [41] we bound all variance values to the range $[0.1, 10]$. For modeling our uncertainties, we consider both ellipsoidal and spherical Gaussian distributions. We estimate the variances of the spherical distributions in two ways: (i) Learn an ellipsoidal distribution and take the average of the variances as the constant spherical variance (spherical-avgpool); or ii) learn a single constant (spherical-one value). Since our probabilistic models include additional parameters, we also report results for the original models with increased capacity (doubling the dimension of their representations in the joint space) to roughly match the capacity of our models.

4.3. Evaluation measures

Following prior work on cross-modal retrieval, we evaluate our models on image-to-text and text-to-image tasks, and report R(ecall)@K (with $K = 1, 5, 10$). R@K is the percentage of queries for which the top-K retrieved samples contains the correct (matching) item. We choose the best models based on the sum of the recall values (rsum) on the validation sets.

Recall measures do not penalize or reward a model based on how good of a match the rest of the top-K retrieved samples are. As explored in [33] with CxC, there are many more plausible matches in a dataset like MS-COCO than the annotated ones. Thus, previous work has proposed R-Precision to complement R@K values [9, 30], calculated as the ratio of all positive items in the top-r retrieved samples for a given query (where r is the number of ground truth matches), averaged over test queries.

We report two versions of the R-Precision, each relying on a different source to determine ground truth matches that are not available in the original image–caption datasets. We report Plausible Match R-Precision (PMRP) on MS-COCO, where we determine the plausible ground truth matches based on the overlap of human-labeled object classes on the MS-COCO images, as in [9]. More concretely, each image and its associated captions are assigned a binary label vector indicating the appearance of a set of object classes in the image. We consider an image–caption pair as a plausible match if their corresponding binary label vector differs at most by a given number of positions, ζ. Following [9], we evaluate on $\zeta \in \{0, 1, 2\}$ and take their average. The drawback of PMRP is that it heavily relies on exhaustive and accurate object annotations, which is difficult to collect. In addition, the assumption that plausible matches can be determined based on the overlap in objects may not always be valid. Specifically, cases can arise where the presence of the same object classes in two images can have different meanings. We thus propose an alternative way to calculating R-Precision, termed RPC^2, that relies on the additionally-collected human annotations in CxC to identify ground-truth plausible matches. Since measuring precision requires special annotations, we can only provide them on MS-COCO, and its extension, CxC. Following previous evaluations on Flickr30k, we only report recall values.

4.4. Ablation study

In the supplemental, we provide an extensive ablation study on the MS-COCO and Flickr30K validation sets over the similarity metric used in our loss, (3), and the form of

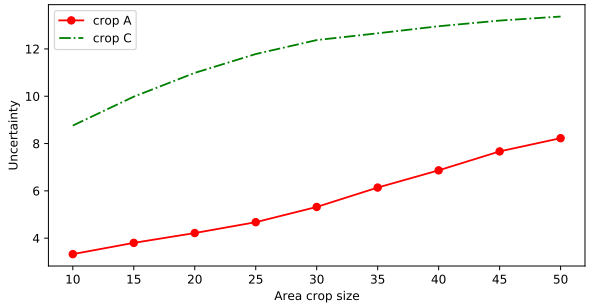
the covariance (i.e., spherical vs. ellipsoidal). Results reported in the remainder of this paper are based on the best combination for each model and dataset, as explained in the next paragraph.

On the MS-COCO dataset, all models perform best with ellipsoidal covariance matrices. We get the best retrieval performance for VSE++ (no fine-tuning) and VSRN with the Wasserstein metric (with 4% difference in rsum between the best-performing and the worst-performing model for VSE++, and 1% difference for VSRN). For the fine-tuned VSE++ (VSE++ft), using the minimum KL metric yields the highest performance (with a 5% difference in rsum). On Flickr30K, the best results for all models are obtained with spherical covariance matrices. The best performance is achieved by using KL divergence for VSE++ (a 2% difference in rsum) and Wasserstein for VSRN (a 3% difference in rsum).

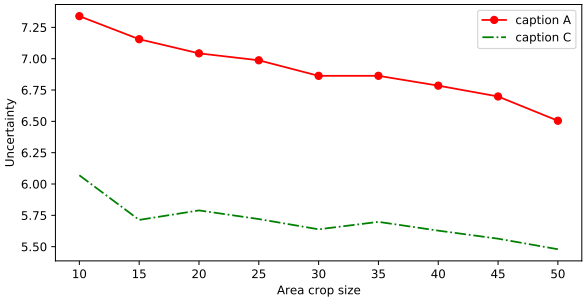
4.5. Retrieval results

Table 1 summarizes the R@K and PMRP results on MS-COCO test pairs. Apart from the results on the original models with and without our proposed probabilistic embeddings, we also include published results for PCME [9] and PVSE [39] for direct comparison. Note, both models propose new architectures and optimize their architecture for cross-modal retrieval. In contrast, our approach can be easily adapted to any architecture with minimal change. Moreover, PCME is the only model that learns probabilistic representations but it does it at the cost of retrieval performance. For PVSE, we report the published recall values in [39] and the published PMRP values in [9]. As can be seen in Table 1, adding probabilistic representations (ours) to both VSE++ and VSRN yields improved (or competitive) performance across the board. As noted in prior work [9, 30, 33], the R-Precision measures are more suited at capturing fine-grained differences in the performance of cross-modal retrieval models. PCME improved on VSRN (state of the art on cross-modal retrieval at the time) when comparing precision (PMRP) but not recall. We can see that adding probabilistic representations to VSRN (VSRN ours) results in further improvements on precision over PCME for both image-to-text and text-to-image retrieval, while substantially outperforming PCME in terms of recall.

As previously mentioned, a more accurate measure of precision is the RPC^2 metric that uses actual human annotations to determine the plausible matches. Table 2 summarizes our finer-grained retrieval results on the MS-COCO test set with the additional CxC annotations. These results are in line with those in Table 1, and further show the superiority of our probabilistic representations. Interestingly, although PCME showed improved precision over all non-probabilistic models (VSE++ with and without finetuning, VSRN, and PVSE) with the PMRP metric, it is only better than VSE++ models based on the more accurate precision measure of RPC^2. Our probabilistic approach added to VSRN achieves the best performance in terms of both precision *and* recall.



(a) Images



(b) Captions

Figure 2. Average uncertainty values of images and captions with varying area thresholds.

Table 3 shows our results on Flickr30K [45]. Note that the R-Precision measures require special annotations that are available for MS-COCO but not for Flickr30K, and as such cannot be reported here. We can see that for VSE++, the probabilistic model (ours) performs better than the base model, both with and without fine-tuning. Compared to the high-capacity model, VSE++ ours performs better for image-to-text and is competitive for text-to-image. For VSRN, we do not see a clear advantage for any of the models. Nonetheless, we stress that our model is competitive and has an associated uncertainty measure that can be used a diagnostic tool.

4.6. Uncertainty and cross-modal ambiguity

Our guiding hypothesis is that the uncertainty encoded in our probabilistic embeddings captures the cross-modal ambiguity in the input. For example, an image with several objects/events can be described in many different ways, each description focusing on a subset of the visual scene. That is, a content-rich image has a high degree of cross-modal ambiguity. In contrast, an image with a single common object/event has a low degree of cross-modal ambiguity. To verify this hypothesis, we create a controlled dataset using the densely-captioned Visual Genome to show that the un-

Model	1K (averaged over five fold)								5K (full test set)							
	image-to-text				text-to-image				image-to-text				text-to-image			
	PMRP	R1	R5	R10	PMRP	R1	R5	R10	PMRP	R1	R5	R10	PMRP	R1	R5	R10
VSE++ *	39.2	58.3	86.1	93.3	39.3	43.7	77.6	87.8	29.3	**35.3**	63.3	75.3	29.1	23.2	50.3	63.5
VSE++ †	39.2	59.0	**87.0**	**94.2**	33.3	44.7	78.4	88.3	29.1	33.8	**64.6**	76.1	28.9	23.3	51.2	64.1
VSE++ ours	**41.1**	**59.3**	86.8	93.8	**41.7**	**45.6**	**79.2**	**89.1**	**31.3**	34.7	**64.6**	**76.6**	**31.2**	**24.5**	**52.4**	**65.6**
VSE++ft [10]	40.6	64.6	90.0	95.7	41.2	52.0	84.3	92.0	29.8	41.3	71.1	81.2	30.0	**30.3**	59.4	72.4
VSE++ft ours	**43.0**	**66.4**	**91.2**	**96.5**	**43.8**	**52.5**	**84.6**	**92.7**	**32.3**	**41.4**	**72.4**	**82.9**	**32.3**	30.0	**60.4**	**72.7**
VSRN *	41.2	**74.0**	94.3	**97.9**	42.4	60.8	88.4	94.1	29.7	50.2	79.6	87.9	29.9	37.9	68.5	79.4
VSRN †	39.5	70.5	93.1	97.1	40.2	58.8	88.2	94.5	27.7	45.3	76.0	86.4	28.1	35.9	66.7	78.3
VSRN ours	**45.8**	73.8	**94.4**	**97.9**	**46.7**	**61.3**	**89.2**	**95.2**	**34.2**	**51.1**	**80.1**	**89.4**	**34.5**	**38.8**	**69.1**	**80.2**
PCME [9]	45.1	68.8	91.6	96.7	46.0	54.6	86.3	93.8	34.1	44.2	73.8	83.6	34.4	31.9	62.1	74.5
PVSE [39]	42.8	69.2	91.6	96.6	43.6	55.2	86.5	93.7	31.8	45.2	74.3	84.5	32.0	32.4	63.0	75.0

Table 1. Retrieval results on MS-COCO. * denotes results generated with the official saved models; † identifies the higher capacity models trained using the official code, and ft stands for fine-tuned. Best results in each group are highlighted in **bold**.

Model	1K (averaged over five fold)								5K (full test set)							
	image-to-text				text-to-image				image-to-text				text-to-image			
	RPC^2	R1	R5	R10	RPC^2	R1	R5	R10	RPC^2	R1	R5	R10	RPC^2	R1	R5	R10
VSE++ *	39.0	59.3	87.1	93.9	40.2	44.7	78.4	88.4	23.8	**37.2**	66.7	78.6	23.4	25.3	54.0	67.3
VSE++ ours	**39.3**	**60.3**	**87.6**	**94.2**	**42.1**	**46.6**	**79.9**	**89.5**	**24.1**	36.7	**68.0**	**80.0**	**24.8**	**26.5**	**55.8**	**69.0**
VSE++ft *	44.1	65.6	90.9	96.1	47.8	53.2	84.9	92.3	**28.5**	43.3	74.2	84.1	**30.1**	**32.5**	62.7	75.3
VSE++ft ours *	**44.2**	**67.6**	**91.9**	**96.8**	**48.0**	**53.6**	**85.2**	**93.1**	28.1	**43.7**	**74.7**	**85.0**	29.9	32.1	**63.4**	**75.5**
VSRN *	50.4	**74.8**	94.8	98.1	54.8	61.8	88.8	94.4	34.3	52.4	81.8	90.0	37.1	40.1	71.1	81.5
VSRN ours	**50.5**	74.5	**94.9**	**98.2**	**55.2**	**62.3**	**89.7**	**95.4**	**34.5**	**53.1**	**82.6**	**91.1**	**37.7**	**40.9**	**71.5**	**82.4**
PCME *	45.9	69.2	92.0	97.0	49.2	55.3	86.5	94.1	29.5	45.2	75.1	85.1	30.7	33.4	64.1	76.3
PVSE *	46.4	70.0	92.2	97.2	50.3	56.4	87.0	94.0	30.3	47.1	77.2	87.0	32.1	34.6	66.0	77.8

Table 2. Retrieval results on the MS-COCO test set with the additional CxC annotations. * denotes results generated using the official saved models and ft stands for fine-tuned. Best results in each group are highlighted in **bold**.

Model	image-to-text			text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
VSE++ [10]	43.7	71.9	82.1	32.3	60.9	72.1
VSE++ †	45.6	73.0	82.1	**34.1**	**62.4**	**73.6**
VSE++ ours	**48.8**	**74.5**	**83.1**	33.3	**62.4**	72.9
VSE++ft [10]	52.9	80.5	87.2	39.6	70.1	79.5
VSE++ft ours	**56.9**	**82.5**	**90.1**	**41.1**	**71.9**	**80.3**
VSRN *	**70.4**	89.2	93.7	**53.0**	77.9	85.7
VSRN †	69.1	**90.3**	**94.4**	52.3	78.5	86.1
VSRN ours	69.2	89.4	94.1	52.3	**79.3**	**86.5**

Table 3. Flickr30k test set retrieval results. * denotes results generated with the official saved models and † the higher capacity models trained using the official code. Best results in each group are highlighted in **bold**.

certainty of an image/caption is indicative of its cross-modal ambiguity. We then investigate the connection between ambiguity and cross-modal retrieval through a controlled binary selection experiment. The rest of this section explains our controlled dataset and experiments in more detail.

Our goal is to compare uncertainties for pairs of images (captions) where one is clearly less ambiguous than the other. To form such pairs, we randomly sample 2000 images from Visual Genome. From each image, we generate a triplet of crops (A, B, C) as follows: For a given area threshold, we select the largest ten crops whose area is less than the threshold. Each of these crops is considered to be fairly unambiguous, since it was created by assigning a tight bounding box around a detected object [21]. We assign the largest among these as the first item in the triplet (crop A). From the remaining nine crops, we choose crop B as the one with the smallest overlap (measured as IoU) with crop A. Crop C is formed by taking the union of crops A and B. Note that this process results in a triplet of differently-sized crops of the same image, formed in a way that ensures crop C contains more objects (and as such is more ambiguous) than crop A. We choose triplets from the same image to control for other visual properties of an image that may contribute to its ambiguity. For corresponding captions we use the Visual Genome annotation for crop A and B. To create a caption for crop C, we concatenate the captions of crops A and B (whose union forms crop C) with the word "and". Fig. 3 shows an example of our cropping process.

To further understand how the connection between cross-modal ambiguity and uncertainty affects cross-modal retrieval, we perform a controlled experiment over the crop triplet dataset that we created as explained above. Following prior work suggesting multiple-choice selection as a finer-grained alternative to retrieval for the evaluation of image–text matching models [13, 14, 37], we frame our experiment as a binary selection task. Specifically, we present each model (either the original VSE++ or VSE++ with probabilistic embeddings) with either caption (crop) A or C as the query, and select between crop (caption) A and C based on their similarities with the query in the joint space.

Fig. 2(a) plots the average uncertainty of crops A and C for different area thresholds ranging from 10% to 50%, measured using our probabilistic VSE++ model. As can be seen, irrespective of the area threshold, the uncertainty of

Figure 3. An example of our dataset preparation for the uncertainty experiments. A and B denote the first and second crop, respectively, and C their union.

model	image-to-text		text-to-image	
	crop A	crop C	caption A	caption C
VSE++*	**59.7**	66.0	**74.8**	54.0
VSE++ ours	58.6	**69.5**	71.7	**58.9**

Table 4. Binary selection accuracy over the 2000 Visual Genome crop triplets.

the union crop (which is expected to contain more objects and be more ambiguous) is always notably higher, as expected. This corresponds to image-to-text binary selection accuracy in Table 4: Looking at the image-to-text columns, we can see that for an unambiguous crop such as A used as the query, the models perform comparably. In contrast, for a more complex and ambiguous crop such as C, our probabilistic model substantially improves upon its point-based embedding counterpart, suggesting that explicitly capturing uncertainty is beneficial.

Fig. 2(b) plots the average uncertainty of captions A and C for the same range of area threshold values as above. Here, we see a different pattern: the uncertainty of the conjoined caption (caption C corresponding to the crop with more objects) is lower than that of caption A. This is expected because the conjoined caption provides more information about the image it can describe, and as such is expected to have a lower degree of cross-modal ambiguity. This corresponds to the text-to-image binary selection accuracy in Table 4: As can be seen in the the text-to-image columns, we do not see an advantage for the probabilistic representations when the task is to select the correct crop given the short caption A as the query. In contrast, for the conjoined caption C, our probabilistic model shows an advantage in selecting the correct crop.

5. Conclusions

In this paper, we proposed a simple yet general and effective approach for learning probabilistic representations for cross-modal retrieval models. We demonstrated the generality of our method by extending extant and distinct point-based cross-modal retrieval models. Through extensive experimentation, we showed that our probabilistic representations yield superior or competitive performance to their point-based analogs and is superior to the lone, concurrent work proposing probabilistic embeddings [9]. Further, our embeddings have the added benefit over point-based ones by providing a measure of uncertainty that we empirically validated to capture the cross-modal ambiguity in images and captions.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 4

[2] Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *TPAMI*, 40(6):1437–1451, 2018. 2

[3] Ben Athiwaratkun and Andrew Gordon Wilson. Multimodal word distributions. In *Annual Meeting of the Association for Computational Linguistics*, pages 1645–1656, 2017. 2

[4] Ben Athiwaratkun and Andrew Gordon Wilson. Hierarchical density order embeddings. In *International Conference on Learning Representations*, 2018. 2

[5] Ben Athiwaratkun, Andrew Gordon Wilson, and Anima Anandkumar. Probabilistic FastText for multi-sense word embeddings. In *Annual Meeting of the Association for Computational Linguistics*, pages 1–11, 2018. 2

[6] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *Machine Learning Research*, 11:1109–1135, 2010. 3

[7] Wei Chen, Yu Liu, Erwin M. Bakker, and Michael S. Lew. Integrating information theory and adversarial learning for cross-modal retrieval. *Pattern Recognition*, 117, 2021. 2

[8] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014. 4

[9] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021. 1, 2, 3, 4, 5, 6, 7

[10] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference*, 2018. 1, 2, 3, 6

[11] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. Devise: A deep visual-semantic embedding model. In *Conference on Neural Information Processing Systems*, pages 2121–2129, 2013. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[13] Micah Hodosh and Julia Hockenmaier. Focused evaluation for image description with binary forced-choice tasks. In *5th ACL Workshop on Vision and Language*, 2016. 3, 6

[14] Hexiang Hu, Ishan Misra, and Laurens van der Maaten. Evaluating text-to-image matching using binary image selection (BISON). In *arXiv preprint arxiv:1901.06595*, 2019. 3, 6

[15] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal LSTM. In *Conference on Computer Vision and Pattern Recognition*, pages 7254–7262, 2017. 2

[16] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311, 2010. 2

[17] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1704–1716, 2012. 2

[18] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 3

[19] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Conference on Neural Information Processing Systems*, pages 5574–5584, 2017. 1

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 4

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2016. 3, 6

[22] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *European Conference on Computer Vision*, pages 201–216, 2018. 2, 3

[23] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *International Conference on Computer Vision*, pages 4653–4661, 2019. 1, 2, 4

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision*, pages 740–755, 2014. 3

[25] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Conference on Neural Information Processing Systems*, pages 13–23, 2019. 2

[27] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Conference on Computer Vision and Pattern Recognition*, pages 10434–10443, 2020. 2

[28] Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes. In *Conference on Neural Information Processing Systems*, pages 5660–5670, 2017. 10

[29] Tanmoy Mukherjee and Timothy Hospedales. Gaussian visual-linguistic embedding for zero-shot recognition. In *Empirical Methods in Natural Language Processing*, pages 912–918, 2016. 2

[30] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision*, pages 681–699, 2020. 4, 5

[31] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Conference on Computer Vision and Pattern Recognition*, pages 2156–2164, 2017. 2

[32] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. In *International Conference on Learning Representations*, 2019. 1, 2

[33] Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 2855–2870, 2021. 3, 4, 5, 10

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems*, pages 8024–8035, 2019. 3

[35] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 2

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 3

[37] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurelie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! Find one mismatch between image and language caption. In *Annual Meeting of the Association for Computational Linguistics*, pages 255–265, 2017. 6

[38] Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. Learning visually-grounded semantics from contrastive adversarial samples. In *Annual Meeting of the Association for Computational Linguistics*, pages 3715–3727, 2018. 2

[39] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019. 2, 3, 5, 6

[40] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations*, 2016. 2

[41] Luke Vilnis and Andrew McCallum. Word representations via Gaussian embedding. In *International Conference on Learning Representations*, 2014. 1, 2, 4

[42] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. CAMP: Cross-modal adaptive message passing for text-image retrieval. In *International Conference on Computer Vision*, pages 5763–5772, 2019. 2

[43] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2020. 2

[44] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. UniVSE: Robust visual semantic embeddings via structured semantic representations. In *Conference on Computer Vision and Pattern Recognition*, pages 6609–6618, 2019. 2

[45] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3, 5

[46] Bowen Zhang, Hexiang Hu, Vihan Jain, Eugene Ie, and Fei Sha. Learning to represent image and text with denotation graph. In *Empirical Methods in Natural Language Processing*, pages 823–839, 2020. 2

[47] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z. Li. Context-aware attention network for image-text retrieval. In *Conference on Computer Vision and Pattern Recognition*, pages 3533–3542, 2020. 2

[48] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2), 2020. 2

6. Appendix

6.1. Similarity metrics

Below, we provide details on the similarity metrics that we use in this work:

1. **Negative Kullback-Liebler (KL) divergence**. This metric is the only asymmetric similarity metric that we use. Since images contain more details than captions (i.e., a caption cannot verbalize all aspects in an image), we expect the captions to have higher uncertainties. Therefore, we take the KL divergence with respect to the caption distribution, i.e., $\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)||\mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c))$, with the following closed form:

$$-\frac{1}{2}\Bigg(\mathrm{tr}(\Sigma_c^{-1}\Sigma_i) + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_c)^{\mathrm{T}}\Sigma_c^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_c) - D + \ln\left(\frac{\det(\Sigma_c)}{\det(\Sigma_i)}\right)\Bigg), \tag{4}$$

where $\mathrm{tr}(\cdot)$ and $\det(\cdot)$ are the trace and determinant matrix operators, respectively. Since we assume diagonal covariance matrices, (4) can be simplified to

$$-\frac{1}{2}\mathrm{tr}\left(\mathrm{diag}\left(\frac{\boldsymbol{\sigma}_i^2}{\boldsymbol{\sigma}_c^2} - \ln\frac{\boldsymbol{\sigma}_i^2}{\boldsymbol{\sigma}_c^2} + \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_c)^2}{\boldsymbol{\sigma}_c^2} - \mathbf{1}\right)\right), \tag{5}$$

where all vector operations are element-wise, $\mathbf{1}$ is a D-dimensional vector of all ones and $\mathrm{diag}(\boldsymbol{a})$ is a diagonal matrix, with $\boldsymbol{a}$ containing the elements on its diagonal. For completeness, we also considered taking the KL divergence with respect to the image embeddings and observed a performance drop in retrieval, as expected. Please see Section 6.2 in this supplemental for more details.

2. **Negative Minimum KL divergence.** We define this measure as a symmetric variant of the KL divergence:

$$-\min\{\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)||\mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c)), \mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c)||\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i))\}. \tag{6}$$

3. **Negative 2-Wasserstein distance**. The Wasserstein distance is a symmetric distance metric between probability distributions. We use the second-order Wasserstein (2-Wasserstein) because it admits the following closed-form for Gaussians [28]:

$$-\sqrt{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_c\|^2 + \mathrm{tr}\left(\Sigma_i + \Sigma_c - 2\left(\Sigma_i^{\frac{1}{2}}\Sigma_c\Sigma_i^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)}, \tag{7}$$

where $\|\cdot\|$ is the Euclidean norm operator. Given our assumption of diagonal covariance matrices, (7) simplifies to

$$-\sqrt{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_c\|^2 + \|\boldsymbol{\sigma}_i^2 - \boldsymbol{\sigma}_c^2\|^2}. \tag{8}$$

6.2. Ablations

We provide ablations on the similarity metrics and distribution shapes on the MS-COCO validation set in Table 5. We estimate the variances of the spherical distributions in two ways: (i) Learn an ellipsoidal distribution and take the average of the variances as the constant spherical variance (spherical-avgpool); or ii) learn a single constant (spherical-one value). As we can see, all models perform best with ellipsoidal distributions. For VSE++ (no fine-tuning) and VSRN, the Wasserstein metric ((8) in the main manuscript) shows the best performance in terms of sum of recalls (rsum), while for VSE++ft the best results are achieved with minimum KL.

Recall that we use the negative of the KL divergence as one of our similarity metrics. Table 5 contains the recall values for the negative KL divergence, when we take the caption distribution as reference, and measure the divergence of the image distributions from this reference distribution. In Table 6, we provide the results (on MS-COCO validation) for the other possibility, where we measure the KL divergence of the caption distributions from the (reference) image distributions. We can see that the sum of recalls (rsum) drops in all cases (with a maximum drop of 2.3%) when we change the order of the distributions for the KL divergence.

6.3. High and low uncertainty images and captions

In this section, we provide examples with high and low uncertainties along with their plausible matches from the CxC dataset [33]. Figs. 4 and 5 show five randomly chosen high-uncertainty images and captions, respectively. Figs. 6 and 7 show five randomly chosen low-uncertainty images and captions, respectively.

6.4. Model stability

Tables 7 and Table 8 provide stability analyses for the two original models, VSE++ and VSRN, and their probabilistic counterparts (ours), on the test portions of MS-COCO and Flickr30K datasets. Specifically, we generate results for five runs of each model and dataset, and report the mean and standard deviation of the rsum values. We can see that for VSE++, the probabilistic model shows consistent improvements over the point-based model on both datasets, with a notably larger mean (and small comparable standard deviation) for VSE++ ours compared to VSE++. For VSRN, which has a more elaborate image encoding and processing pipeline, we observe a larger variation in performance between runs. Nonetheless, based on these numbers

model	metric	shape	image-to-text			text-to-image			rsum
			R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ ours	KL (4)	spherical-avgpool	57.2	86.5	94.0	43.9	78.1	88.9	448.6
VSE++ ours	KL (4)	spherical-one value	56.2	86.5	94.1	43.5	77.7	88.8	446.8
VSE++ ours	KL (4)	ellipsoidal	61.3	88.1	94.1	46.6	80.6	90.4	461.1
VSE++ ours	minKL (6)	spherical-avgpool	54.7	87.2	94.7	43.7	77.9	88.7	446.9
VSE++ ours	minKL (6)	spherical-one value	57.1	87.6	93.8	43.8	77.7	88.8	448.8
VSE++ ours	minKL (6)	ellipsoidal	60.0	87.6	95.2	46.8	80.9	91.0	446.9
VSE++ ours	Wasserstein (7)	spherical-avgpool	58.1	88.3	94.3	44.9	79.8	89.3	454.7
VSE++ ours	Wasserstein (7)	spherical-one value	59.3	87.5	94.4	45.0	79.4	90.1	455.7
VSE++ ours	**Wasserstein (7)**	ellipsoidal	61.3	88.3	95.2	47.1	81.0	91.0	463.9
VSE++ft ours	KL (4)	spherical-avgpool	68.8	92.4	97.3	53.7	86.4	93.7	492.3
VSE++ft ours	KL (4)	spherical-one value	61.7	90.1	95.7	49.3	82.9	92.3	472.0
VSE++ft ours	KL (4)	ellipsoidal	70.0	92.8	96.3	54.3	86.8	93.7	493.9
VSE++ft ours	minKL (6)	spherical-avgpool	62.4	90.5	96.1	50.0	84.4	92.6	476.0
VSE++ft ours	minKL (6)	spherical-one value	61.4	90.1	95.6	49.1	83.1	91.9	471.2
VSE++ft ours	**minKL (6)**	ellipsoidal	68.8	93.2	97.6	55.4	86.9	93.5	495.4
VSE++ft ours	Wasserstein (7)	spherical-avgpool	61.8	90.3	95.5	52.8	85.2	93.1	478.7
VSE++ft ours	Wasserstein (7)	spherical-one value	67.9	92.6	97.1	54.2	86.3	93.8	491.9
VSE++ft ours	Wasserstein (7)	ellipsoidal	67.3	93.1	97.4	54.2	86.7	93.8	492.5
VSRN ours	KL (4)	spherical-avgpool	76.5	95.9	98.5	62.2	90.0	96.0	519.1
VSRN ours	KL (4)	spherical-one value	77.5	96.7	98.7	62.0	89.9	95.7	520.5
VSRN ours	KL (4)	ellipsoidal	76.5	96.3	98.1	62.9	91.0	95.9	520.7
VSRN ours	minKL (6)	spherical-avgpool	76.0	96.0	98.3	63.5	90.9	96.4	521.1
VSRN ours	minKL (6)	spherical-one value	75.3	96.2	98.5	60.8	90.7	96.2	517.7
VSRN ours	minKL (6)	ellipsoidal	77.1	96.2	98.8	63.6	90.5	96.2	522.4
VSRN ours	Wasserstein (7)	spherical-avgpool	76.8	96.8	99.2	63.2	90.5	96.0	522.5
VSRN ours	Wasserstein (7)	spherical-one value	76.4	96.5	98.9	63.1	91.2	95.9	522.0
VSRN ours	**Wasserstein (7)**	ellipsoidal	77.3	97.1	98.7	63.7	91.2	96.0	524.0

Table 5. Retrieval ablation study on the MS-COCO validation set. ft stands for fine-tuned. Best results in terms of sum of recalls in each group are highlighted in **bold**. The numbers in parantheses denote the equation number of the similarity metric in the main manuscript.

model	shape	image-to-text			text-to-image			rsum
		R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ ours	spherical	52.6	85.5	93.0	43.1	76.6	87.5	438.3
VSE++ ours	ellipsoidal	57.5	87.6	94.0	45.6	79.6	89.9	454.3
VSRN ours	spherical	75.6	95.1	98.4	62.2	89.3	95.5	516.1
VSRN ours	ellipsoidal	75.5	96.7	98.8	62.4	90.7	96.3	520.4

Table 6. Retrieval results on MS-COCO validation set with the negative of KL divergence similarity between the image and caption distributions taken with respect to image representations.

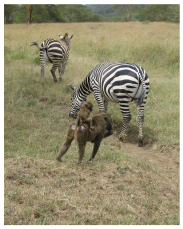
model	MS-COCO		Flickr30K
	1K	5K	
VSE++	447.74 ± 0.51	310.06 ± 0.81	363.42 ± 2.02
VSE++ ours	452.70 ± 0.50	317.02 ± 1.07	372.48 ± 1.86

Table 7. VSE++ rsum means $\pm$ standard deviations.

we can still see that the probabilistic VSRN (ours) is competitive (if not better than) the point-based VSRN.

model	MS-COCO		Flickr30K
	1K	5K	
VSRN	507.68 ± 0.80	401.66 ± 1.40	466.48 ± 1.88
VSRN ours	510.76 ± 2.02	405.02 ± 2.82	468.96 ± 2.64

Table 8. VSRN rsum means $\pm$ standard deviations.



Two zebras and two monkeys walking on the grass.
Two giraffes and another animal are on green grass.
A baboon and two zebras grazing on the savannah.
A baboon and its baby eat by two zebras in the grass.
Monkey standing behind two zebras as they graze.

A herd of elephants standing around in the middle of a pen.
The elephant family is walking near the rocks.
Two full grown elephants and one baby elephant outside.
A bunch of elephants that are in a line.
Two elephants and a baby are walking by the rock.

A herd of elephants walking through a shallow river.
There is a herd of elephants standing together in the water.
Baby elephants are standing by adult elephants in water.
A large group of elephants gathered in water.
A large group of elephants in a lake.

A bird standing in the wooded area with leaves all around.
A photo of a red jungle fowl standing on leaves.
A chicken standing in the dry brown leaves.
This wild bird is walking on twigs and leaves.
A red and brown bird walking through brush of the same color.

A close shot of a cow standing in the wild.
The large cow is standing in a field of tall grass and flowers.
A white cow stares straight ahead while standing in high grass.
A cow that is standing in the grass.
A big white cow, yellow flowers, the blue sky.

Figure 4. High uncertainty images and their corresponding plausible matches.

Two different brands but similar looking devices sitting by each other.

A small coyote is seen in the back of some tall grass.

A couple of animals walking through the grass in a forest.

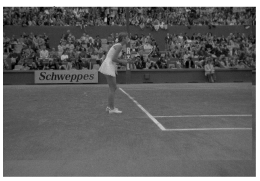
Two bears kiss each other on the nose.

A pregnant zebra walking through the tall grass.

Figure 5. High uncertainty captions and their corresponding plausible matches.

Two skiers in the snow out in the open.
Two skiers travel past a vast field of blowing snow.
A couple of men riding skis on a snow covered slope.
Two people dressed in snow gear with skis and poles walking through the snow.
Two people with backpacks skiing on snow covered slopes.
Two skiers standing near a sign in the snow.
Snowboards and skis in the snow with two people in the background.
A person in walking on skies in a snow covered mountain.
Several cross-country skiers on a snowy path in the woods.

A man on a beach flying a kite.
A person flying a red kite on a beach.
A man flying a red kite in a blue sky.
A person is holding the end of a red kite's string.
Person walking on a beach carrying a camera bag.
People out at the beach flying kites.
A person is flying a kite on a beach.
A woman flies a kite against a blue clear sky.
A man near a body of water flies a kite.



The woman is playing tennis while a crowd watches.
A woman waiting for a serve while playing tennis.
A young woman playing tennis in a tournament.
A female tennis player stand on a tennis court.
A woman leaning over a bit near the end of a court.
A black woman playing tennis runs and swings at the ball.
A woman in white shirt and skirt playing a game of tennis.
A girl that is stretched out on a tennis court.
A woman hitting a tennis ball on a tennis court during a match.

Woman in sportswear playing tennis on a tennis court.
A woman leans in playing on a tennis court.
A woman reaching out to swing a tennis racket.
A young woman is on a blue tennis court with arms outstretched.
A girl that is stretched out on a tennis court.
A girl reaches for a ball with her tennis racket.
A lady on a tennis court swinging her racket.
A woman standing on a tennis court with a racket in her hand.
A beautiful woman in white and pink holding a tennis racquet.

A person riding a skateboard up the side of a wall.
A skateboarder is in mid air with the board up against a wall.
A young person attempts to skateboard along the wall.
A person riding a skateboard on a wall.
A male skateboarder is doing a skateboard trick.
A young man riding a skateboard over a stone bench.
A person riding a skate board at a skate park.
A skateboarder rides along a diagonally slanted wall.

Figure 6. Low uncertainty images and their corresponding plausible matches.

Figure 7. Low uncertainty captions and their corresponding plausible matches.