Probabilistic Embeddings for Cross-Modal Retrieval – Supplementary Materials –

Supplementary Materials

We include additional materials in this document. We describe additional details on PCME to complement the main paper (§A). Various probabilistic distances are introduced (§B). We provide the experimental protocol details (§C), ablation studies (§D), and additional results (§E). Finally, more uncertainty analyses are shown (§F).

A. More details for PCME

In this section, we provide details for PCME.

A.1. The uniformity loss

Recently, Wang *et al*. [17] proposed the uniformity loss which enforces the feature vectors to distribute uniformly on the unit hypersphere. In Wang *et al*. [17], the uniformity loss was shown to lead to better representations for L2 normalized features. Since our μ vectors are projected to the unit L2 hypersphere, we also employ the uniformity loss to learn better representations. We apply the uniformity loss on the joint embeddings $\mathcal{Z} = \{v_1^1, t_1^1, \ldots, v_B^J, t_B^J\}$ in the mini-batch size of B as follows:

$$\mathcal{L}_{\text{Unif}} = \sum_{z,z' \in \mathcal{Z} \times \mathcal{Z}} e^{-2\|z-z'\|_2^2}. \tag{A.1}$$

A.2. Connection between the soft contrastive loss and the MIL objective of PVSE

In the main text, we presented an analysis based on gradients to study how the loss function in Equation (1) handles plurality in cross-modal matches and learns uncertainties in data. Here we make connections with the MIL loss used by PVSE (§3.1.1, [16]); this section follows the corresponding section in the main paper.

To build connections with PVSE, consider a one-hot weight array $w_{jj'}$ where, given that (v, t) is a positive pair, the "one" value is taken only by the single pair (j, j') whose distance is smallest. Define $w_{jj'}$ for a negative pair (v, t) conversely. Then, we recover the MIL loss used in PVSE, where only the best match among J^2 predictions are utilized. As we see in the experiments, our *softmax* weight scheme provides more interpretable and performant supervision for the uncertainty than the *argmax* version used by PVSE.

B. Probabilistic distances

We introduce probabilistic distance variants to measure the distance between two normal distributions $p = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q = \mathcal{N}(\mu_2, \sigma_2^2)$. All distance functions are non-negative and become zero if and only if two distributions are identical. Extension to multivariate Gaussian distributions with diagonal variance can be simply derived by taking the summation over the dimension-wise distances.

Kullback–Leibler (KL) divergence measures the difference between two distributions as follows:

$$\begin{aligned} KL(p, q) &= \int \log \frac{p}{q} dp \\ &= \frac{1}{2}\left[\log \frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2}\right]. \end{aligned} \tag{B.1}$$

KL divergence is not a metric because it is asymmetric ($KL(p, q) \neq KL(q, p)$) and does not satisfy the triangular inequality. If q has a very small variance, nearly zero, the KL divergence between p and q will be explored. In other words, if we have a very certain embedding, which has nearly zero variance, in our gallery set, then the certain embedding will be hardly retrieved by KL divergence measure. In the latter section, we will show that KL divergence leads to bad retrieval performances in the real-world scenario.

Jensen-Shannon (JS) divergence is the average of forward ($KL(p, q)$) and reverse ($KL(q, p)$) KL divergences. Unlike KL divergence, the square root of JS divergence is a metric function.

$$JS(p, q) = \frac{1}{2}\left[KL(p, q) + KL(q, p)\right]. \tag{B.2}$$

Like KL divergence, JS divergence still has division term by variances σ_1, σ_2, it can be numerically unstable when the variances are very small.

Probability product kernels [9] are generalized inner product for two distributions, that is:

$$PPK(p, q) = \int p(z)^\rho q(z)^\rho dz. \tag{B.3}$$

When $\rho = 1$, it is called the expected likelihood kernel (ELK), and when $\rho = 1/2$, it is called Bhattacharyya's affinity [1], or Bhattacharyya kernel.

Expected likelihood kernel (ELK) is a special case of PPK when $\rho = 1$ in Equation (B.3). In practice, we take log to compute ELK as follows:

$$ELK(p,q) = \frac{1}{2}\left[\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2} + \log(\sigma_1^2+\sigma_2^2)\right]. \quad \text{(B.4)}$$

Bhattacharyya kernel (BK) is another special case of PPK when $\rho = 1/2$ in Equation (B.3). The log BK is defined as follows:

$$BK(p,q) = \frac{1}{4}\left[\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2} + 2\log(\frac{\sigma_2}{\sigma_1}+\frac{\sigma_1}{\sigma_2})\right]. \quad \text{(B.5)}$$

Wasserstein distance is a metric function of two distributions on a given metric space M. The Wasserstein distance between two normal distributions on $\mathbb{R}^1$, 2-Wasserstein distance, is defined as follows:

$$W(p,q)^2 = (\mu_1-\mu_2)^2 + \sigma_1 - \sigma_2{}^2. \quad \text{(B.6)}$$

C. Experimental Protocol Details

We introduce the cross-modal retrieval benchmarks considered in this work. We discuss the issues with the current practice for the evaluation and introduce new alternatives.

C.1. Plausible Match R-Precision (PMRP) details

In this work, we seek more reliable sources of pairwise similarity measurements through class and attribute labels on images. For example, on the CUB caption dataset, we have established the positivity of pairs by the criterion that a pair (i, c) is positive if and only if both elements in the pair belong to the same bird class. Similarly, on the COCO caption dataset, we judge the positivity through the multiple class labels (80 classes total) attached per image: a pair (i, c) is positive if and only if the binary class vectors for the two instances, $y^i, y^c \in \{0,1\}^{80}$, differ at most at ζ positions (Hamming Distance). In MS-COCO 5k test images, 48 images do not have instance labels; we omit them during the evaluation. Note that because we use R-Precision, the ratio of positive items in top-r retrieved items where r is the number of the ground-truth matches, increasing ζ will make r larger, and will penalize methods more, which retrieve irrelevant items.

In Figure C.1, we visualize the number of distinct categories per image in the MS-COCO validation set. In the figure, we can observe that about the half of the images have more than two categories. To avoid penalty caused by almost neglectable objects (as shown in Figure C.2), we set $\zeta = 2$ for measuring the PMRP score. For PMRP with different ζ rather than 2, results can be found in §E.

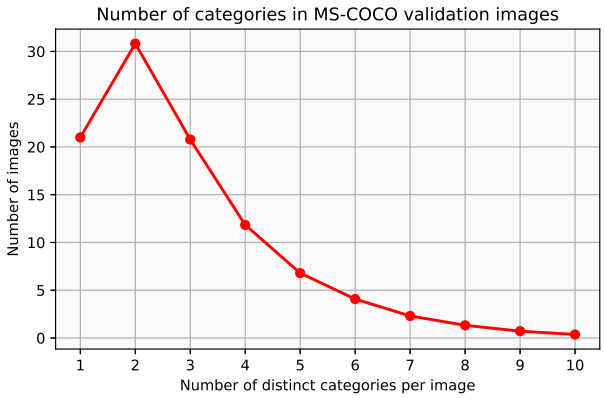


Figure C.1. **Number of distinct categories in MS-COCO validation set.** Images that have more than 10 categories are omitted.

C.2. Implementation details

Common. As in Faghri *et al*. [6], we use ResNet [7] pre-trained on ImageNet and the pre-trained GloVe with 2.2M vocabulary [14] for initializing the visual and textual encoders ($f_\mathcal{V}$, $f_\mathcal{T}$). We first warm-up the models by training the head modules for each modality, with frozen feature extractors. Afterwards, the whole parameters are fine-tuned in an end-to-end fashion. We use the ResNet-152 backbone with embedding dimension $D = 1024$ for MS-COCO and ResNet-50 with $D = 512$ for CUB. For all experiments, we set the number of samples $J = 7$ (the detailed study is in §E). We use AdamP optimizer [8] with the cosine learning rate scheduler [12] for stable training.

MS-COCO. We follow the evaluation protocol of [10] where the validation set is added to the training pool (referred to as rV in [5, 6]). Our training and validation splits contain 113,287 and 5,000 images, respectively. We report results on both 5K and (the average over 5-fold) 1K test sets.

Hyperparameter search protocol. We validate the initial learning rate, number of epochs for the warm-up and fine-tuning, and other hyperparameters on the 150 CUB training classes and the MS-COCO caption validation split. For MS-COCO, we use the initial learning rate as 0.0002, 30 warm-up and 30 finetune epochs. Weights for regularizers $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{Unif}}$ are set to 0.00001 and 0, respectively. For CUB Caption, the initial learning rate is 0.0001, the number of warm-up epochs 10 and fine-tuning epochs 50. Weights for regularizers $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{Unif}}$ are set to 0.001 and 10, respectively. For both datasets, models are always trained with Cutout [4] and random caption dropping [2] augmentation strategies with 0.2 and 0.1 erasing ratios, respectively. The initial values for a, b in Equation (3) are set to -15 and 15 for COCO (-5 and 5 for CUB), respectively.

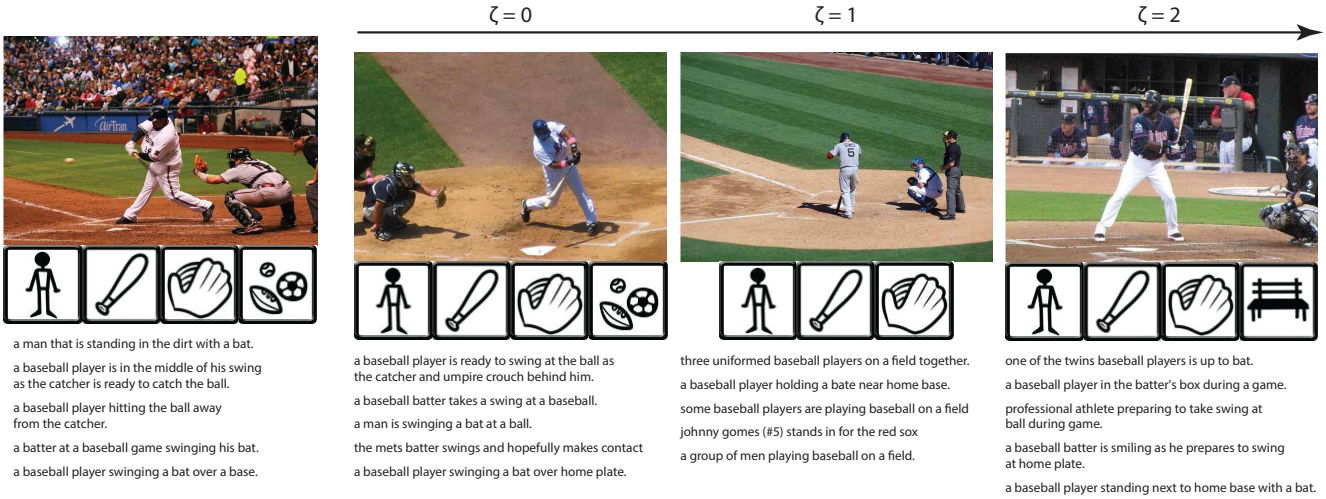


Figure C.2. **MS-COCO plausible match examples.** The plausible examples of the most left instance from $\zeta = 0$ to $\zeta = 2$. The contained instance classes, ζ, figure and captions are shown.

C.3. CUB 2D toy experiment details

We select nine bird classes from CUB caption; three swimming birds ("Western Grebe", "Pied Billed Grebe", "Pacific Loon"), three small birds ("Vermilion Flycatcher", "Black And White Warbler", "American Redstart"), and three woodpeckers ("Red Headed Woodpecker", "Red Bellied Woodpecker", "Downy Woodpecker").

We slightly modify PCME to learn 2-dimensional embeddings. For the image encoder, we use the same structure as the other experiments, but omitting the attention modules from the μ and σ modules. For the caption encoder, we train 1024-dimensional bi-GRU on top of GloVe vectors and apply two 2D projections to get the 1024 dimensional μ and σ embedding. The other training details are the same as the other CUB caption experiments.

D. Ablation studies

We provide ablation studies on PCME for regularization terms, σ module architectures, the number of samples J during training, and embedding dimension D.

Regularizing uncertainty. PCME predicts probabilistic outputs. We have considered uncertainty-specific regularization strategy in the main paper, the information bottleneck loss $\mathcal{L}_{\text{KL}}$ and the uniform loss $\mathcal{L}_{\text{Unif}}$. We study the benefits of those ingredients. Table D.1 shows our results. We report cross-validated MAP@R [13] on the 150 class training CUB caption datasets. The KL loss increases the sigma values to a meaningful range (from $e^{-13.01} \approx 2.2 \times 10^{-6}$ to $e^{-3.84} \approx 0.02$. The uniformity loss prevents the uncertainty from collapsing and slightly improves performances.

$\mathcal{L}_{\text{KL}}$	$\mathcal{L}_{\text{Unif}}$	i2t MAP@R	t2i MAP@R	Image $\mathbb{E}[\log \sigma]$	Caption $\mathbb{E}[\log \sigma]$
✗	✗	10.56	13.32	-13.01	-8.77
✓	✗	10.57	13.77	-3.84	-3.89
✗	✓	10.56	13.31	-11.26	-7.59
✓	✓	**10.65**	**13.84**	-3.63	-3.64

Table D.1. **Regularization for uncertainty.** Cross-validated MAP@R performances on CUB training set, with and without KL and uniformity loss terms. The scale estimate $\mathbb{E}[\log \sigma]$ is an averaged value over the σ dimensions as well as the validation samples.

Method	DoF(σ)	i2t	t2i
PCME μ only	0	24.7	25.6
PCME isotropic	1	25.7	26.0
PCME	512	**26.3**	**26.8**

Table D.2. **DoF for σ.** R-Precision on the CUB Caption test set.

DoF for σ. Though by default we parametrize the full diagonal elements of the covariance matrix $\Sigma \in \mathbb{R}^{D \times D}$ with the vector $\sigma \in \mathbb{R}^{D}$, one may parametrize σ more cheaply via *e.g.* a scalar, by restricting the embedding distribution family to isotropic Gaussians. Table D.2 shows the trade-off between the degree of freedom (DoF) for σ and the R-Precision of PCME. Indeed, allowing greater degrees of freedom for σ brings better performance. Figure D.1 shows the average variance values for each dimension, which supports that the learned variances require high DoF.

Architecture study. Table D.3 shows the architecture design comparisons for PCME on CUB Caption test split. In the table, applying local attention to both μ and σ modules performs the best. Furthermore, we ablate sigmoid and `LN`

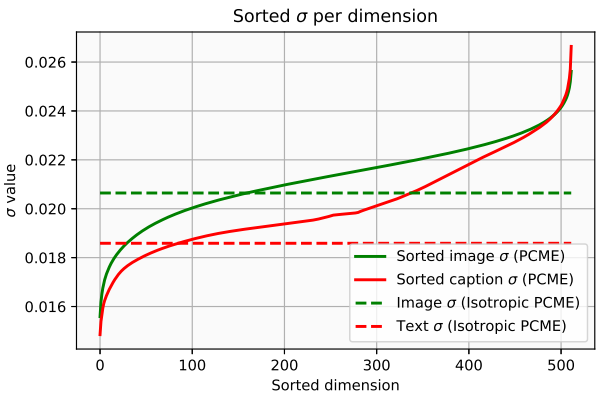


Figure D.1. **How isotropic are variances?** Sorted values of variance are compared against the trained values of isotropic PCME. Results on CUB test set.

μ local attention	σ local attention	I-to-T R-Precision	T-to-I R-Precision
✗	✗	25.60	25.85
✗	✓	24.65	25.15
✓	✗	25.01	25.52
✓	✓	**26.28**	**26.77**

$s(\cdot)$ & LN in σ module	I-to-T R-Precision	T-to-I R-Precision
✓	23.81	24.58
✗	**26.28**	**26.77**

Table D.3. **Architectures for μ and σ.** Architecture design choices comparison on CUB caption test split.

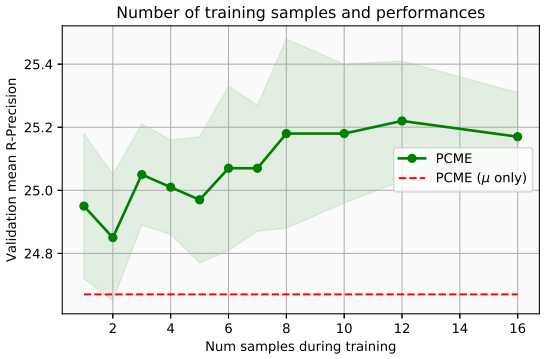


Figure D.2. **Number of samples.** The cross-validated PCME performances against the number of samples J during training.

parts of σ modules, which can restrict the representation of variances. As a result, limiting representations by sigmoid and layer norm harms the final performances.

Number of samples during training. In Figure D.2, we report the cross-validated mean R-Precision scores by varying the number of samples J during training. In the figure, we observe that larger J leads to higher performances. In practice, we choose $J = 7$ for computation budgets.

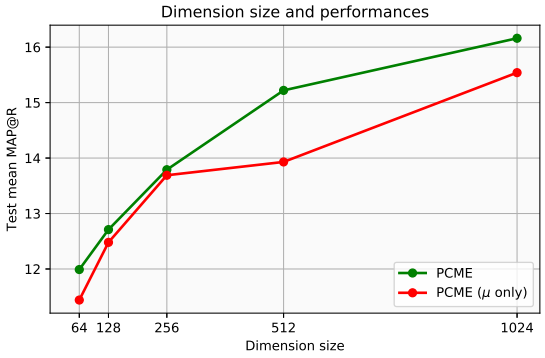


Figure D.3. **Embedding dimensions.** The PCME performance against the embedding dimensions D.

PCME variant	Sampling	Test-time Similarity Metric	Space complexity	i2t R-P	t2i R-P
μ only	✗	Mean only	$O(N)$	24.70	25.64
PCME	✗	Mean only	$O(N)$	26.14	26.67
	✗	KL-divergence	$O(2N)$	21.99	20.92
	✗	JS-divergence	$O(2N)$	25.06	25.55
	✗	ELK	$O(2N)$	25.33	25.87
	✗	Bhattacharyya	$O(2N)$	24.93	25.27
	✗	2-Wasserstein	$O(2N)$	<u>26.16</u>	<u>26.69</u>
	✓	Average L2	$O(J^2N)$	26.11	26.64
	✓	Match prob	$O(J^2N)$	**26.28**	**26.77**

Table E.1. **Pairwise distances for distributions.** There are many options for computing the distance between two distributions. What are the space complexity and retrieval performances for each option? R-P stands for the R-Precision.

Embedding dimensions. Performances against different embedding space dimensions for PCME μ only and PCME are illustrated in Figure D.3. In all embedding dimensions, our stochastic approach (PCME) consistently outperforms the deterministic approach (PCME μ only).

E. More results

In this section, we provide additional experimental results for PCME on CUB Caption and COCO Caption.

E.1. More results on similarity measures for retrieval at test time

In Table E.1, we report the full retrieval results obtained by the different distribution distances discussed in §B. As discussed in §B, KL-divergence even shows worse results than the "Mean only" baseline, a non-probabilistic distance.

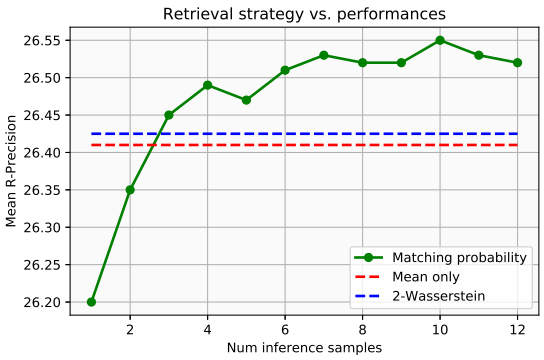


Figure E.1. **Comparison of different retrieval strategies.**

We also report the performances against the number of samples of matching probability in Figure E.1. In the figure, the matching probability strategy shows better results than non-sampling strategies from $J = 3$, and larger J leads to better performances. Due to the computation complexity, we use $J = 7$ in Table E.1.

E.2. Discussion on hardest negative mining

Since Recall@K is widely used for the evaluation of many cross-modal retrieval tasks, many recent cross-modal retrieval methods optimize Recall@1 directly by the hardest negative mining (HNM) strategy [6], that is:

$$\begin{aligned} &\max_{t'}[\alpha + \text{sim}(v, t') - \text{sim}(v, t)] \\ &+ \max_{v'}[\alpha + \text{sim}(v', t) - \text{sim}(v, t)], \end{aligned} \tag{E.1}$$

where sim is the cosine similarity. This strategy neglects all other possible positive candidates, but only considers the most similar positive and negative pairs. To reveal that HM strategy disadvantages to learn the global structure, we measure two metrics on CUB caption, R-Precision and Recall@1. For non-HM strategy, we replace max to $\sum$ in Equation (E.1). Figure E.2 shows R-Precision and recall@1 performances with different mining strategies. In the figure, PVSE with HNM strategy shows higher Recall@1 by increasing the number of embeddings K (36.3 → 37.6 → 41.1), but at the same time, it reduces the R-Precision scores (21.4 → 20.4 → 19.2). On the other hand, for all K, Non-HNM strategy PVSE results show worse R@1 than HNM results but achieves higher R-Precision performances. In Table 3, we show that this phenomenon is also observed in MS-COCO by measuring PMRP scores.

E.3. Full results for CUB and COCO

CUB Caption. We report the full results on CUB Caption test data for unseen 50 classes and seen 150 classes in Table E.2 and Table E.3, respectively. In both splits, PCME shows the best R-Precision performances against baselines.

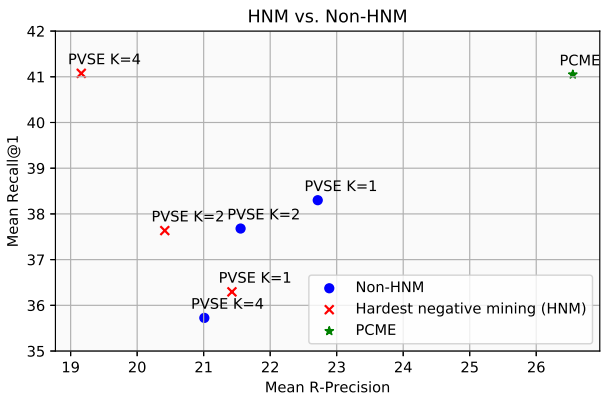


Figure E.2. **Hardest negative mining (HNM) vs. Non-HNM.**

Method	HNM	Image-to-text R-P	Image-to-text R@1	Text-to-image R-P	Text-to-image R@1
VSE0	✗	22.35	44.19	22.57	32.71
PVSE K=1	✗	22.65	43.11	22.78	33.49
PVSE K=2	✗	21.62	44.05	21.49	31.31
PVSE K=4	✗	21.12	40.51	20.90	30.94
PVSE K=1	✓	22.34	40.88	20.51	31.71
PVSE K=2	✓	19.67	47.29	21.16	27.98
PVSE K=4	✓	18.38	**47.76**	19.94	34.39
PCME μ only	✗	24.70	46.38	25.64	**35.50**
PCME	✗	**26.28**	46.92	**26.77**	35.22

Table E.2. **Comparison on CUB Caption unseen 50 class test set.** R-P and R@1 stand for R-Precision and Recall@1, respectively. The usage of the hardest negative mining (HNM) is indicated.

Method	HNM	Image-to-text R-P	Image-to-text R@1	Text-to-image R-P	Text-to-image R@1
VSE0	✗	19.85	40.88	18.72	25.51
PVSE K=1	✗	19.69	40.65	18.72	25.58
PVSE K=2	✗	18.84	41.45	17.72	24.99
PVSE K=4	✗	18.31	38.08	17.21	23.54
PVSE K=1	✓	18.98	38.77	18.23	23.49
PVSE K=2	✓	17.62	44.24	17.71	22.78
PVSE K=4	✓	17.47	**44.98**	17.44	26.19
PCME μ only	✗	20.65	42.70	20.16	**26.94**
PCME	✗	**20.87**	43.10	**20.37**	26.47

Table E.3. **Comparison on CUB Caption seen 150 class test set.** R-P and R@1 stand for R-Precision and Recall@1, respectively. The usage of the hardest negative mining (HNM) is indicated.

COCO Caption. We report the full results on MS-COCO Caption 1k test images and 5k test images in Table E.4 and Table E.5, respectively. We also report additional experiments on PVSE such as larger K ($K = 4$), a different

Method	D	Image-to-text				Text-to-image			
		PMRP	R@1	R@5	R@10	PMRP	R@1	R@5	R@10
VSE++ BMVC'18 [6]	1024	-	64.6	90.0	95.7	-	52.0	84.3	92.0
PVSE K=1 CVPR'19 [16]	1024	40.3*	66.7	91.0	96.2	41.9*	53.5	85.1	92.7
PVSE K=2 CVPR'19 [16]	1024 × 2	42.8*	69.2	91.6	96.6	43.7*	55.2	86.5	93.7
PVSE K=4 CVPR'19 [16]	1024 × 4	41.5	68.0	91.9	96.6	42.7	54.1	85.5	92.9
PVSE K=1 + SHM [15]	1024 × 1	41.6	66.1	91.4	96.4	42.4	53.6	85.5	93.0
PVSE K=2 + SHM [15]	1024 × 2	39.0	65.1	90.9	96.5	39.4	53.1	85.4	93.0
VSRN ICCV'19 [11]	2048	41.2*	76.2	94.8	98.2	42.4*	62.8	89.7	95.1
VSRN + AOQ ECCV'20 [3]	2048 × 2	44.7*	**77.5**	**95.5**	**98.6**	45.6*	**63.5**	**90.5**	**95.8**
PCMEμ only	1024	45.0	68.0	92.0	96.2	45.9	54.6	86.3	93.8
PCME	1024 × 2	**45.1**	68.8	91.6	96.7	**46.0**	54.6	86.3	93.8

Table E.4. **1K MS-COCO results.** Plausible Match R-Precision (PMRP), Recall@K results on MS-COCO 1k test images. "∗" denotes results produced by the published models.

Method	D	Image-to-text				Text-to-image			
		PMRP	R@1	R@5	R@10	PMRP	R@1	R@5	R@10
VSE++ BMVC'18 [6]	1024	-	41.3	71.1	81.2	-	30.3	59.4	72.4
PVSE K=1 CVPR'19 [16]	1024	29.3*	41.7	73.0	83.0	30.1*	30.6	61.4	73.6
PVSE K=2 CVPR'19 [16]	1024 × 2	31.8*	45.2	74.3	84.5	32.0*	32.4	63.0	75.0
PVSE K=4 CVPR'19 [16]	1024 × 4	30.5	43.0	72.8	83.6	31.0	31.2	61.5	74.4
PVSE K=1 + SHM [15]	1024 × 1	30.6	41.1	71.6	82.7	30.8	30.9	60.8	73.7
PVSE K=2 + SHM [15]	1024 × 2	28.1	40.7	70.8	81.9	27.8	29.9	60.4	73.4
VSRN ICCV'19 [11]	2048	29.7*	53.0	81.1	89.4	29.9*	40.5	70.6	81.1
VSRN + AOQ ECCV'20 [3]	2048 × 2	33.0*	**55.1**	**83.3**	**90.8**	33.5*	**41.1**	**71.5**	**82.0**
PCMEμ only	1024	34.0	43.5	73.1	84.2	34.3	31.7	62.2	74.9
PCME	1024 × 2	**34.1**	44.2	73.8	83.6	**34.4**	31.9	62.1	74.5

Table E.5. **Comparison on 5K MS-COCO.** Plausible Match R-Precision (PMRP), Recall@K results on MS-COCO 5k test images. "∗" denotes results produced by the published models.

negative mining strategy (semi-hard negative mining [15]. In the tables, although PCME shows slightly worse R@1 results than PVSE K=2, PCME outperforms PVSE K=2 in PMRP scores.

Also, we report PMRP scores of four methods (PVSE [16], VSRN [11], VSRN + AOQ [3] and PCME) by varying ζ for PMRP in Figure E.3. In the figure, PMRP scores for VSRN and VSRN + AOQ are getting worse by increasing ζ, in other words, theses method shows less coherence if we allow one missing or altering object class in the retrieved items. On the other hand, PCME shows even increased performance with $\zeta > 0$, in other words, PCME retrieves more plausible items than other methods.

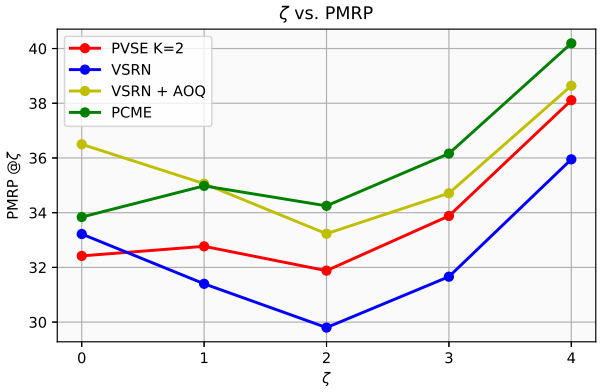


Figure E.3. **PMRP by varying ζ.** Plausible Match R-Precision scores for four methods with $\zeta = \{0, 1, 2\}$.

F. More uncertainty analysis

Uncertainty estimation by PCME brings interesting insights for the cross-modal retrieval tasks. In this section, we show additional uncertainty analysis for PCME.

F.1. Corruption vs. uncertainty in MS-COCO

As Figure 7, we illustrate the uncertainty level by varying corruption levels on pixels and words in Figure F.1. The

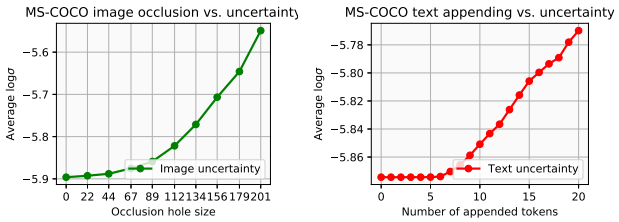


Figure F.1. σ **captures ambiguity in COCO Caption.** Average $\log \sigma$ values at different ratios of erased pixels (for images) and appended <unk> tokens (for captions).

left figure shows the uncertainty levels against occluded pixels. As we expected, more occlusion leads to higher uncertainty. The right figure shows the uncertainty levels against the number of appended <unk> tokens.

F.2. Frequent words for each uncertainty bin

Figure F.2 shows the frequent words per each uncertainty bin. We use term frequency–inverse document frequency (TF-IDF) as the frequent counter, defined as follows:

$$\text{TF-IDF}(i) = (1 + \log n_i) \log \frac{N}{n_i}, \tag{F.1}$$

where N is the number of total captions, and n_i is the number of captions which contain word i. For the image word frequency, we use their ground truth captions for computing TF-IDF scores.

F.3. Example uncertain samples

We visualize the uncertain images and captions, and their corresponding retrieved items in Figure F.3 and Figure F.4. Interestingly, the retrieved captions and images are plausible results for the given query items. These qualitative results also show how the Recall@1 measure is noisy, and the proposed Plausible Match R-Precision (PMRP) is a more plausible and reliable measure to compare different retrieval methods.

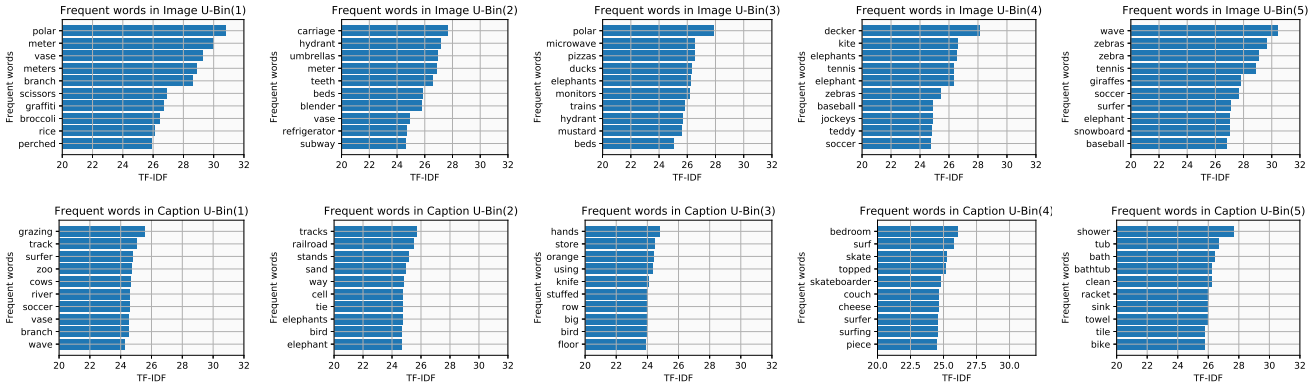
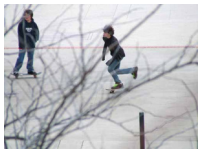


Figure F.2. **Frequent words in each uncertainty bin.** Term frequency–inverse document frequency (TF-IDF) sorted word frequencies are shown for each uncertainty bin (U-Bin, ascending order) for image (upper row) and caption (bottom row) modalities.

Query Image (σ=0.0052)



Retrieved captions

A boy riding on ski's down a slope.
A young boy is attempting to slide down a slope.
A kid is riding down the street on a skateboard.
a man with warm clothes skating on the snow
a young person riding skis on a snowy field
a person skating in very much snow with warm clothes

GT captions

Two boys riding skateboards in the street, behind tree branches.
two young people riding skate boards on a flat surface
Two young men riding skateboards across a parking lot.
two young men skateboarding in an open area during winter
A couple of kids riding on top of skateboards.

Query image (σ = 0.0054)

Retrieved captions

Two people in the midst of a tennis match on a grass court.
Two men on grass court playing a game of tennis.
Two men playing doubles tennis on a grass court.
Two men playing tennis on a grass field.
a couple of people play a game of tennis on a grass surface
A male tennis players on the court with rackets.

GT captions

A couple of men holding tennis racquets on a tennis court.
Two men playing tennis at a somewhat large facility.
Two men playing doubles tennis on a grass court.
A couple of tennis players during a couples game about to deliver a hit.
A male tennis players on the court with rackets.

Query image (σ = 0.0054)

Retrieved captions

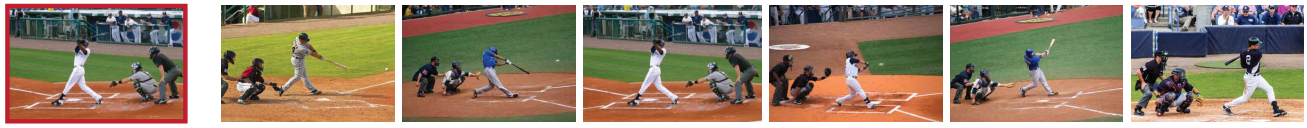
A surfer riding a wave in a blue ocean.
A wet suited surfer riding the crest of an azure wave
a male surfing a large ocean wave on a white surfboard
The surfer is working on riding the big wave.
A surfer is riding on a large wave.
A surf boarder who is riding a wave.

GT captions

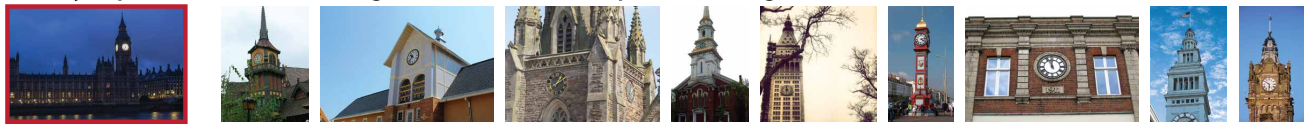
A surfer is on his board in the middle of an ocean spraying wave.
A man on a surfboard riding a wave
A man is surfing a small wave in the ocean.
A man riding on a wave on a surf board.
a person riding a surf board on a wave

Figure F.3. **Uncertain image examples.** Highly uncertain images, retrieved captions by PCME, and their ground truth captions are shown.

Query caption (σ = 0.0046): A batter is swinging at a ball at the game.



Query caption (σ = 0.0046): a large clock tower is on top of a building.



Query caption (σ = 0.0047): A man playing tennis outside during a sunny day.

GT image **Retrieved images**

Figure F.4. **Uncertain caption examples.** Highly uncertain captions, retrieved images by PCME, and their ground truth image are shown.

References

[1] Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943. 2

[2] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proc. CoNLL*, pages 10–21, 2016. 2

[3] Tianlang Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. In *Proc. ECCV*, 2020. 6

[4] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[5] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proc. CVPR*, 2018. 2

[6] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proc. BMVC*, 2018. 2, 5, 6

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 2

[8] Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Gyuwan Kim, Youngjung Uh, and Jung-Woo Ha. Adamp: Slowing down the slowdown for momentum optimizers on scale-invariant weights. In *Proc. ICLR*, 2021. 2

[9] Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5(Jul):819–844, 2004. 1

[10] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, pages 3128–3137, 2015. 2

[11] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proc. ICCV*, pages 4654–4662, 2019. 6

[12] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *Proc. ICLR*, 2017. 2

[13] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Proc. ECCV*, 2020. 3

[14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proc. EMNLP*, 2014. 2

[15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. CVPR*, pages 815–823, 2015. 6

[16] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proc. CVPR*, pages 1979–1988, 2019. 1, 6

[17] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proc. ICML*, 2020. 1