Probabilistic Embeddings for Cross-Modal Retrieval

Sanghyuk Chun[1] Seong Joon Oh[1] Rafael Sampaio de Rezende[2] Yannis Kalantidis[2] Diane Larlus[2]

[1]NAVER AI Lab [2]NAVER LABS Europe

Abstract

Cross-modal retrieval methods build a common representation space for samples from multiple modalities, typically from the vision and the language domains. For images and their captions, the multiplicity of the correspondences makes the task particularly challenging. Given an image (respectively a caption), there are multiple captions (respectively images) that equally make sense. In this paper, we argue that deterministic functions are not sufficiently powerful to capture such one-to-many correspondences. Instead, we propose to use Probabilistic Cross-Modal Embedding (PCME), where samples from the different modalities are represented as probabilistic distributions in the common embedding space. Since common benchmarks such as COCO suffer from non-exhaustive annotations for cross-modal matches, we propose to additionally evaluate retrieval on the CUB dataset, a smaller yet clean database where all possible image-caption pairs are annotated. We extensively ablate PCME and demonstrate that it not only improves the retrieval performance over its deterministic counterpart but also provides uncertainty estimates that render the embeddings more interpretable. Code is available at https://github.com/naver-ai/pcme.

1. Introduction

Given a query and a database from different modalities, cross-modal retrieval is the task of retrieving the database items which are most relevant to the query. Most research on this topic has focused on the image and text modalities [5, 9, 25, 51, 58]. Typically, methods estimate embedding functions that map visual and textual inputs into a common embedding space, such that the cross-modal retrieval task boils down to the familiar nearest neighbour retrieval task in a Euclidean space [9, 51].

Building a common representation space for multiple modalities is challenging. Consider an image with a group of people on a platform preparing to board a train (Figure 1). There is more than one possible caption describing this image. "People waiting to board a train in a train platform" and "The metro train has pulled into a large station" were two of the choices from the COCO [5] annotators. Thus, the common representation has to deal with the fact that an image potentially matches with a number of different captions. Conversely, given a caption, there may be multiple manifestations of the caption in visual forms. The multiplicity of correspondences across image-text pairs stems in part from the different natures of the modalities. All the different components of a visual scene are thoroughly and passively captured in a photograph, while language descriptions are the product of conscious choices of the key relevant concepts to report from a scene. All in all, a common representation space for image and text modalities is required to model the one-to-many mappings in both directions.

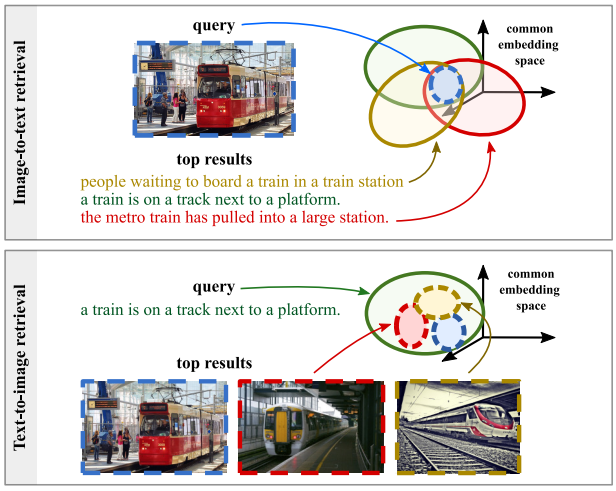


Figure 1. We propose to use **probabilistic embeddings** to represent images and their captions as probability distributions in a common embedding space suited for cross-modal retrieval. These distributions gracefully model the uncertainty which results from the multiplicity of concepts appearing in a visual scene and implicitly perform many-to-many matching between those concepts.

Standard approaches which rely on vanilla functions do not meet this necessary condition: they can only quantify one-to-one relationships [9, 51]. There have been attempts to introduce multiplicity. For example, Song and Soleymani [45] have introduced Polysemous Visual-Semantic

Embeddings (PVSE) by letting an embedding function propose K candidate representations for a given input. PVSE has been shown to successfully capture the multiplicity in the matching task and to improve over the baseline built upon one-to-one functions. Others [25] have computed region embeddings obtained with a pre-trained object detector, establishing multiple region-word matches. This strategy has led to significant performance gains at the expense of a significant increase in computational cost.

In this work, we propose **Probabilistic Cross-Modal Embedding (PCME)**. We argue that probabilistic mapping is an effective representation tool that does not require an explicit many-to-many representation as is done by detection-based approaches, and further offers a number of advantages. First, PCME yields uncertainty estimates that lead to useful applications like estimating the difficulty or chance of failure for a query. Second, the probabilistic representation leads to a richer embedding space where set algebras make sense, whereas deterministic ones can only represent similarity relations. Third, PCME is complementary to the deterministic retrieval systems.

As harmful as the assumption of one-to-one correspondence is for the method, the same assumption has introduced confusion in the evaluation benchmarks. For example, MS-COCO [5] suffers from non-exhaustive annotations for cross-modal matches. The best solution would be to explicitly and manually annotate all image-caption pairs for evaluation. Unfortunately, this process does not scale, especially for a large-scale dataset like COCO. Instead, we propose a smaller yet cleaner cross-modal retrieval benchmark using CUB [55] and more sensible evaluation metrics.

Our contributions are as follows. (1) We propose Probabilistic Cross-Modal Embedding (PCME) to properly represent the one-to-many relationships in joint embedding spaces for cross-modal retrieval. (2) We identify shortcomings with existing cross-modal retrieval benchmarks and propose alternative solutions. (3) We analyse the joint embedding space using the uncertainty estimates provided by PCME and show how intuitive properties arise.

2. Related work

Cross-modal retrieval. In this work, we are interested in image and text cross-modal retrieval. Much research is dedicated to learning a metric space that jointly embeds images and sentences [8, 9, 10, 18, 25, 45, 47]. Early works [11, 23] relied on Canonical Correlation Analysis (CCA) [13] to build joint embedding spaces. Frome *et al*. [10] use a hinge rank loss for triplets built from both modalities. Wang *et al*. [51] expand on this idea by also training on uni-modal triplets to preserve the structure inherent to each modality in the joint space. Faghri *et al*. [9] propose to learn such space with a triplet loss, and only sample the hardest negative with respect to a query-positive pair.

One of the drawbacks of relying on a single global representation is its inability to represent the diversity of semantic concepts present in an image or in a caption. Prior work [16, 54] observed a split between *one-to-one* and *many-to-many* matching in visual-semantic embedding spaces characterized by the use of one or several embedding representations per image or caption. Song and Soleymani [45] build many global representations for each image or sentence by using a multi-head self-attention on local descriptors. Other methods use region-level and word-level descriptors to build a global image-to-text similarity from many-to-many matching. Li *et al*. [25] employ a graphical convolutional network [22] for semantic reasoning of region proposals obtained from a Faster-RCNN [40] detector. Veit *et al*. [49] propose a conditional embedding approach to solve the multiplicity of hashtags, but it does not rely on a joint embedding space, hence cannot be directly applied to cross-modal retrieval.

Recently, the most successful way of addressing many-to-many image-to-sentence matching is through joint visual and textual reasoning modules appended on top of separate region-level encoders [24, 28, 30, 31, 34, 53, 54, 60]. Most of such methods involve cross-modal attention networks and report state-of-the-art results on cross-modal retrieval. This, however, comes with a large increase in computational cost at test time: pairs formed by the query and every database entry need to go through the reasoning module. Focusing on scalability, we choose to build on top of approaches that directly utilize the joint embedding space and are compatible with large-scale indexing.

Finally, concurrent to our work, Wray *et al*. [56] consider cross-modal video retrieval and discusses similar limitations of the one-to-one correspondence assumptions for evaluation. They propose to consider semantic similarity proxies computed on captions for a more reliable evaluation on standard video retrieval datasets.

Probabilistic embedding. Probabilistic representations of data have a long history in machine learning [32]. They were introduced in 2014 for word embeddings [50], as they gracefully handle the inherent hierarchies in language, since then, a line of research has explored different distribution families for word representations [26, 35, 36]. Recently, probabilistic embeddings have been introduced for vision tasks. Oh *et al*. [37] proposed the Hedged Instance Embedding (HIB) to handle the one-to-many correspondences for metric learning, while other works apply probabilistic embeddings to face understanding [43, 3], 2D-to-3D pose estimation [46], speaker diarization [44], and prototype embeddings [42]. Our work extends HIB to joint embeddings between images and captions, in order to represent the different levels of granularities in the two domains and to implicitly capture the resulting one-to-many associations. Recently Schönnfeld *et al*. [41] utilized Variational Autoen-

coders [20] for zero-shot recognition. Their latent space is conceptually similar to ours, but is learned and used in very different ways: they simply use a 2-Wasserstein distance as their distribution alignment loss and learn classifiers on top, while PCME uses a probabilistic *contrastive* loss that enables us to use the latent features directly for retrieval. To our knowledge, PCME is the first work that uses probabilistic embeddings for multi-modal retrieval.

3. Method

In this section, we present our **Probabilistic Cross-Modal Embedding** (**PCME**) framework and discuss its conceptual workings and advantages.

We first define the cross-modal retrieval task. Let $\mathcal{D} = (\mathcal{C}, \mathcal{I})$ denote a vision and language dataset, where $\mathcal{I}$ is a set of images and $\mathcal{C}$ a set of captions. The two sets are connected via ground-truth matches. For a caption $c \in \mathcal{C}$ (respectively an image $i \in \mathcal{I}$), the set of corresponding images (respectively captions) is given by $\tau(c) \subseteq \mathcal{I}$ (respectively $\tau(i) \subseteq \mathcal{C}$). Note that for every query q, there may be multiple cross-modal matches ($|\tau(q)| > 1$). Handling this multiplicity will be the central focus of our study.

Cross-modal retrieval methods typically learn an embedding space $\mathbb{R}^D$ such that we can quantify the subjective notion of "similarity" into the distance between two vectors. For this, two embedding functions $f_{\mathcal{V}}$, $f_{\mathcal{T}}$ are learned to map image and text samples into the common space $\mathbb{R}^D$.

3.1. Building blocks for PCME

We introduce two key ingredients for PCME: joint visual-textual embeddings and probabilistic embeddings.

3.1.1 Joint visual-textual embeddings

We describe how we learn visual and textual encoders. We then present a previous attempt at addressing the multiplicity of cross-modal associations.

Visual encoder $f_{\mathcal{V}}$. We use the ResNet image encoder [14]. Let $z_v = g_{\mathcal{V}}(i) : \mathcal{I} \to \mathbb{R}^{h \times w \times d_v}$ denote the output before the global average pooling (GAP) layer. Visual embedding is computed via $v = h_{\mathcal{V}}(z_v) \in \mathbb{R}^D$ where in the simplest case $h_{\mathcal{V}}$ is the GAP followed by a linear layer. We modify $h_{\mathcal{V}}$ to let it predict a distribution, rather than a point.

Textual encoder $f_{\mathcal{T}}$. Given a caption c, we build the array of word-level descriptors $z_t = g_{\mathcal{T}}(c) \in \mathbb{R}^{L(c) \times d_t}$, where $L(c)$ is the number of words in c. We use the pre-trained GloVe [38]. The sentence-level feature t is given by a bidirectional GRU [6]: $t = h_{\mathcal{T}}(z_t)$ on top of the GloVe features.

Losses used in prior work. The joint embeddings are often learned with a contrastive or triplet loss [9, 10].

Polysemous visual-semantic embeddings (PVSE) [45] are designed to model one-to-many matches for cross-modal retrieval. PVSE adopts a multi-head attention block on top of the visual and textual features to encode K possible embeddings per modality. For the visual case, each visual embedding $v^k \in \mathbb{R}^D$ for $k \in \{1, \dots, K\}$ is given by: $v^k = \texttt{LN}\left(h_{\mathcal{V}}(z_v) + s(w^1 \texttt{att}^k_{\mathcal{V}}(z_v) z_v)\right)$, where $w^1 \in \mathbb{R}^{d_v \times D}$ are the weights of fully connected layers, s is the sigmoid function and $\texttt{LN}$ is the LayerNorm [1]. $\texttt{att}^k_{\mathcal{V}}$ denotes the k-th attention head of the visual self-attention $\texttt{att}_{\mathcal{V}}$. Textual embeddings t^k for $k \in \{1, \dots, K\}$ are given symmetrically by the multi-head attention: $t^k = \texttt{LN}\left(h_{\mathcal{T}}(z_t) + s(w^2 \texttt{att}^k_{\mathcal{C}}(z_t) z_t)\right)$. PVSE learns the visual and textual encoders with the multiple instance learning (MIL) objective, where only the best pair among the K^2 possible visual-textual embedding pairs is supervised.

3.1.2 Probabilistic embeddings for a single modality

Our PCME models each sample as a distribution. It builds on the Hedged Instance Embeddings (HIB) [37], a single-modality methodology developed for representing instances as a distribution. HIB is the probabilistic analogue of the contrastive loss [12]. HIB trains a probabilistic mapping $p_\theta(z|x)$ that not only preserves the pairwise semantic similarities but also represents the inherent uncertainty in data. We describe the key components of HIB here.

Soft contrastive loss. To train $p_\theta(z|x)$ to capture pairwise similarities, HIB formulates a soft version of the contrastive loss [12] widely used for training deep metric embeddings. For a pair of samples (x_α, x_β), the loss is defined as:

$$\mathcal{L}_{\alpha\beta}(\theta) = \begin{cases} -\log p_\theta(m|x_\alpha, x_\beta) & \text{if } \alpha, \beta \text{ is a match} \\ -\log\left(1 - p_\theta(m|x_\alpha, x_\beta)\right) & \text{otherwise} \end{cases} \tag{1}$$

where $p_\theta(m|x_\alpha, x_\beta)$ is the *match probability*.

Factorizing match probability. [37] has factorized $p_\theta(m|x_\alpha, x_\beta)$ into the match probability based on the embeddings $p(m|z_\alpha, z_\beta)$ and the encoders $p_\theta(z|x)$. This is done via Monte-Carlo estimation:

$$p_\theta(m|x_\alpha, x_\beta) \approx \frac{1}{J^2} \sum_{j}^{J} \sum_{j'}^{J} p(m|z^j_\alpha, z^{j'}_\beta) \tag{2}$$

where z^j are samples from the embedding distribution $p_\theta(z|x)$. For the gradient to flow, the embedding distribution should be reparametrization-trick-friendly [21].

Match probability from Euclidean distances. We compute the sample-wise match probability as follows:

$$p(m|z_\alpha, z_\beta) = s(-a\|z_\alpha - z_\beta\|_2 + b) \tag{3}$$

where (a, b) are learnable scalars and $s(\cdot)$ is sigmoid.

3.2. Probabilistic cross-modal embedding (PCME)

We describe how we learn a joint embedding space that allows for probabilistic representation with PCME.

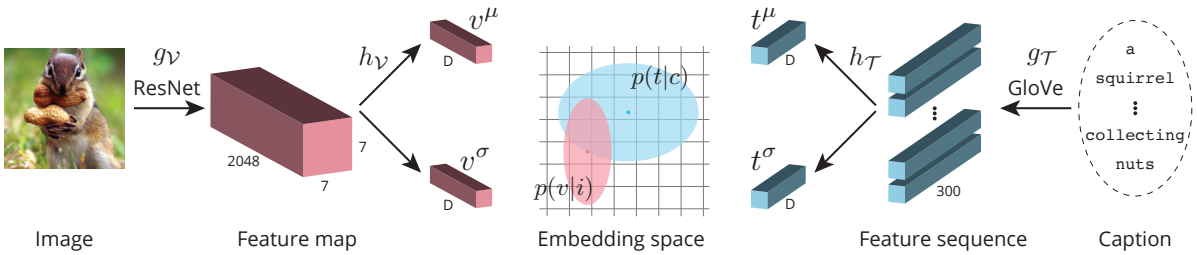


Figure 2. **Method overview.** The visual and textual encoders for Probabilistic Cross-Modal Embedding (PCME) are shown. Each modality outputs mean and variance vectors in $\mathbb{R}^D$, which represent a normal distribution in $\mathbb{R}^D$.

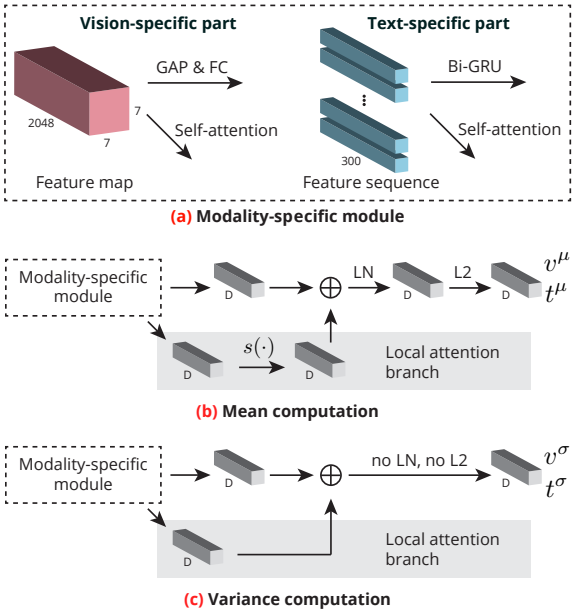


Figure 3. **Head modules.** The visual and textual heads ($h_\mathcal{V}$, $h_\mathcal{T}$) share the same structure, except for modality-specific modules (a). The mean (b) and variance (c) computations differ: variance module does not involve sigmoid $s(\cdot)$, LayerNorm (LN), and L2 projection.

3.2.1 Model architecture

An overview of PCME is shown in Figure 2. PCME represents an image i and caption c as normal distributions, $p(v|i)$ and $p(t|c)$ respectively, over the same embedding space $\mathbb{R}^D$. We parametrize the normal distributions with mean vectors and diagonal covariance matrices in $\mathbb{R}^D$:

$$\begin{aligned} p(v|i) &\sim N\left(h^\mu_\mathcal{V}(z_v), \text{diag}(h^\sigma_\mathcal{V}(z_v))\right) \\ p(t|c) &\sim N\left(h^\mu_\mathcal{T}(z_t), \text{diag}(h^\sigma_\mathcal{T}(z_t))\right) \end{aligned} \tag{4}$$

where $z_v = g_\mathcal{V}(i)$ is the feature map and $z_t = g_\mathcal{T}(c)$ is the feature sequence (§3.1.1). For each modality, two head modules, h^μ and h^σ, compute the mean and variance vectors, respectively. They are described next.

Local attention branch. Inspired by the PVSE architecture (§3.1.1, [45]), we consider appending a *local attention branch* in the head modules (h^μ, h^σ) both for image and caption encoders. See Figure 3 for the specifics. The local attention branch consists of a self-attention based aggregation of spatial features, followed by a linear layer with a sigmoid activation function. We will show with ablative studies that the additional branch helps aggregating spatial features more effectively, leading to improved performance.

Module for μ versus σ. Figure 3 shows the head modules h^μ and h^σ, respectively. For $h^\mu_\mathcal{V}$ and $h^\mu_\mathcal{T}$, we apply sigmoid in the local attention branch and add the residual output. In turn, LayerNorm (LN) [1] and L2 projection operations are applied [45, 48]. For $h^\sigma_\mathcal{V}$ and $h^\sigma_\mathcal{T}$, we observe that the sigmoid and LN operations overly restrict the representation, resulting in poor uncertainty estimations (discussed in §D). We thus do not use sigmoid, LN, and L2 projection for the uncertainty modules.

Soft cross-modal contrastive loss. Learning the joint probabilistic embedding is to learn the parameters for the mappings $p(v|i) = p_{\theta_v}(v|i)$ and $p(t|c) = p_{\theta_t}(t|c)$. We adopt the probabilistic embedding loss in Equation (1), where the match probabilities are now based on the cross-modal pairs (i, c): $\mathcal{L}_{\text{emb}}(\theta_v, \theta_t; i, c)$, where $\theta = (\theta_v, \theta_t)$ are parameters for visual and textual encoders, respectively. The match probability is now defined upon the visual and textual features: $p_\theta(m|i, c) \approx \frac{1}{J^2}\sum_j^J \sum_{j'}^J s(-a\|v^j - t^{j'}\|_2 + b)$ where v^j and $t^{j'}$ follow the distribution in Equation (4).

Additional regularization techniques. We consider two additional loss functions to regularize the learned uncertainty. Following [37], we prevent the learned variances from collapsing to zero by introducing the KL divergence loss between the learned distributions and the standard normal $\mathcal{N}(0, I)$. We also employ the *uniformity loss* that was recently introduced in [52], computed between all embeddings in the minibatch. See §A.1 for more details.

Sampling SGD mini-batch. We start by sampling B ground-truth image-caption matching pairs $(i, c) \in \mathcal{G}$. Within the sampled subset, we consider *every* positive and negative pair dictated by the ground truth matches. This would amount to B matching pairs and $B(B-1)$ non-matching pairs in our mini-batch.

Measuring instance-wise uncertainty. The covariance matrix predicted for each input represents the inherent uncertainty for the data. For a scalar uncertainty measure, we

take the determinant of the covariance matrix, or equivalently the geometric mean of the σ's. Intuitively, this measures the volume of the distribution.

3.2.2 How does our loss handle multiplicity, really?

We perform a gradient analysis to study how our loss in Equation (1) handles multiplicity in cross-modal matches and learn uncertainties in data. In §A.2, we further make connections with the MIL loss used by PVSE (§3.1.1, [45]).

We first define the distance logit: $l_{jj'} := -a\|v^j - t^{j'}\|_2 + b$ and compare the amount of supervision with different (j, j') values. To see this, take the gradient on $l_{jj'}$.

$$\frac{\partial \mathcal{L}_{\text{emb}}}{\partial l_{jj'}} = \begin{cases} w_{jj'} \cdot (1 - s(l_{jj'})) & \text{for positive match} \\ -w_{jj'} \cdot s(l_{jj'}) & \text{for negative match} \end{cases} \tag{5}$$

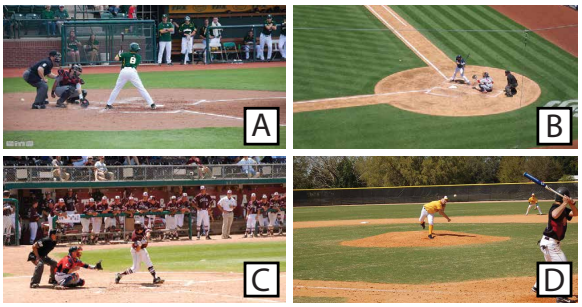
$$w_{jj'} := \frac{e^{\pm l_{jj'}}}{\sum_{\alpha\alpha'} e^{\pm l_{\alpha\alpha'}}} \quad \text{where } \pm \text{ is the positivity of match.}$$

We first observe that if $w_{jj'} = 1$, then Equation (5) is exactly the supervision from the soft contrastive loss (Equation (1)). Thus, it is the term $w_{jj'}$ that let the model learn multiplicity and represent associated uncertainty.

To study the behavior of $w_{jj'}$, first assume that (v, t) is a positive pair. Then, $w_{jj'}$ is the softmax over the pairwise logits $l_{jj'}$. Thus, pairs with smaller distances $\|v^j - t^{j'}\|_2$ have greater weights $w_{jj'}$ than distant ones. Similarly, if (v, t) is negative pair, then $w_{jj'}$ assigns greater weights on distant pairs than close ones. In other words, $w_{jj'}$ gives more weights on pair samples that correctly predicts the distance relationships on the embedding space. This results in a reward structure where wrong similarity predictions do not get penalized significantly, as long as there is at least one correct similarity prediction. Such a reward encourages the embeddings to produce more diverse samples and hedge the bets through non-zero values of σ predictions.

3.2.3 Test-time variants

Unlike methods that employ cross-modal reasoning modules [24, 28, 30, 31, 34, 53, 54, 60], computing match probabilities at test time for PCME reduces to computing a function over pairwise Euclidean distances. This means that the probabilistic embeddings of PCME can be used in various ways for computing the match probabilities at test time, with different variants having different computational complexities. The options are split into two groups. **(i) Sampling-based variants.** Similar to training, one can use Monte-Carlo sampling (Equation (2)) to approximate match probabilities. Assuming J samples, this requires $O(J^2)$ distance computations per match, as well as $O(J^2)$ space for every database entry. This implies that J plays an important role in terms of test time complexity. **(ii) Non-sampling variants.** One can simply use the distances based on μ to approximate match probabilities. In this case, both time and space complexities become $O(1)$. We ablate this variant ("μ only") in our experiments, as it is directly comparable to deterministic approaches. We also may use any distributional distance measures with closed-form expressions for Gaussian distributions. Examples include the 2-Wasserstein distance, Jensen Shanon (JS) divergence, and Expected Likelihood Kernel (ELK). We ablate them as well. The details of each probabilistic distance can be found in §B.



a) A baseball player swinging a bat at a ball.
b) A baseball player is getting ready to hit a ball.
c) A baseball player standing next to home plate holding a bat.
d) A group of baseball players at the pitch.

Figure 4. Can you match the captions to the images? In the COCO annotations, each of the four captions corresponds to (only) one of the four images (Answer: p:D, a:C, c:B, q:A).

4. Experiments

We present experimental results for PCME. We start with the experimental protocol and a discussion on the problems with current cross-modal retrieval benchmarks and evaluation metrics, followed by alternative solutions (§4.1). We then report experimental results on the CUB cross-modal retrieval task (§4.2) and COCO (§4.3). We present an analysis of the embedding space in §4.4.

4.1. Experimental protocol

We use ResNet [14] pre-trained on ImageNet and the pre-trained GloVe with 2.2M vocabulary [38] for initializing the visual and textual encoders. Training proceeds in two phases: a warm-up phase where only the head modules are trained, followed by end-to-end fine-tuning of all parameters. We use a ResNet-152 (resp. ResNet-50) backbone with embedding dimension $D = 1024$ (resp. $D = 512$) for MS-COCO (resp. CUB). For both datasets, models are always trained with Cutout [7] and random caption dropping [2] augmentation strategies with 0.2 and 0.1 erasing ratios, respectively. We use the AdamP optimizer [15] with the cosine learning rate scheduler [29] for stable training. More implementation details are provided in §C.2. Hyperparameter details and ablations are presented in §D.

4.1.1 Metrics for cross-modal retrieval

Researchers have long been aware of many potentially positive matches in the cross-modal retrieval evaluation sets. They use metrics that reflect such consideration.

Many works report the **Recall@k** (R@k) metrics with varying numbers for k. This evaluation policy, with larger values of k, becomes more lenient to plausible wrong predictions prevalent in COCO. However, it achieves leniency at the cost of failing to penalize obviously wrong retrieved samples. The lack of penalties for wrongly retrieved top-k samples may be complemented by the precision metrics.

Musgrave *et al*. [33] proposed the **R-Precision** (R-P) metric as an alternative; for every query q, we compute the ratio of positive items in the top-r retrieved items, where $r = |\tau(q)|$ is the number of ground-truth matches. This precision metric has a desirable property that a retrieval model achieves the perfect R-Precision score if and only if it retrieves all the positive items before the negatives.

For R-Precision to make sense, all the existing positive pairs in a dataset must be annotated. Hence, we expand the existing ground truth matches by seeking further plausible positive matches in a database through extra information (*e.g*. class labels for COCO). More concretely, a pair (i, c) is declared positive if the binary label vectors for the two instances, $y^i, y^c \in \{0, 1\}^{d_{label}}$, differ at most at ζ positions. In practice, we consider multiple criteria $\zeta \in \{0, 1, 2\}$ and average the results with those ζ values. We refer to metrics based on such class-based similarity as **Plausible Match (PM)** because we incentivize models to retrieve plausible items. We refer to the R-Precision metric based on the Plausible Match policy as **PMRP**. More details in §C.1.

4.1.2 Cross-modal retrieval benchmarks

COCO Captions [5] is a widely-used dataset for cross-modal retrieval models. It consists of 123,287 images from MS-COCO [27] with 5 human-annotated captions per image. We present experimental results on COCO. We follow the evaluation protocol of [17] where the COCO validation set is added to the training pool (referred to as rV or rVal in [8, 9]). 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.

The problem with COCO as a cross-modal retrieval benchmark is the binary relevance assignment of image-caption pairs (i, c). As a result, the number of matching captions $\tau(i)$ for an image i is always 5. Conversely, the number of matching images $\tau(c)$ for a caption c is always 1. All other pairs are considered non-matching, independent of semantic similarity. This is far from representing the semantic richness of the dataset. See Figure 4 for an illustration. While all 4×4 possible pairs are plausible positive pairs, 12 of them are assigned negative labels during training and evaluation. This results in noisy training and, more seriously, unreliable evaluation results.

We re-purpose the CUB 200-2011 [55] as a more reliable surrogate for evaluating cross-modal retrieval models. We utilize the caption annotations by Reed *et al*. [39]; they consist of ten captions per image on CUB images (11,788 images of 200 fine-grained bird categories). False positives are suppressed by the fact that the captions and images are largely homogeneous within a class. False negatives are unlikely to happen because the images contain different types of birds across classes and the captions are generated under the instruction that the annotators should focus on class-distinguishing characteristics [39].

We follow the class splits proposed by Xian *et al*. [57], where 150 classes are used for training and validation, and the remaining 50 classes are used for the test. The hyperparameters are validated on the 150 training classes. We refer to this benchmark as *CUB Captions*.

4.2. Results on CUB

Similarity measures for retrieval at test time. We have discussed alternative similarity metrics that PCME may adopt at test time (§ 3.2.3). The "Mean only" metric only uses the h^μ features, as in deterministic retrieval scenarios. It only requires $O(N)$ space to store the database features. Probabilistic distance measures like ELK, JS-divergence, and 2-Wasserstein, require the storage for μ and σ features, resulting in the doubled storage requirement. Sampling-based distance computations, such as the average L2 distance and match probability, need J^2 times the storage required by the Mean-only baseline.

We compare the above variants in Table 1 and §E.1. First of all, we observe that PCME, with any test-time similarity measure, mostly improves over the deterministically trained PCME (μ-only training). Even if the test-time similarity is computed as if the embeddings are deterministic (Mean only), PCME training improves the retrieval performances (24.7% to 26.1% for i2t and 25.6% to 26.7% for t2i). Other cheaper variants of probabilistic distances, such as 2-Wasserstein, also result in reasonable performances (26.2% and 26.7% for i2t and t2i, respectively), while introducing only twice the original space consumption. The best performance is indeed attained by the similarity measure using the match probability, with 26.3% and 26.8% i2t and t2i performances, respectively. There exists a trade-off between computational cost and performance and the deterministic test-time similarity measures. We use the match probability measure at test time for the rest of the paper.

Comparison against other methods. We compare PCME against VSE0 [9] and PVSE [45] in Table 2. As an important ingredient for PVSE, we consider the use of the hardest negative mining (HNM). We first observe that

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
	✗	ELK	$O(2N)$	25.33	25.87
	✗	JS-divergence	$O(2N)$	25.06	25.55
	✗	2-Wasserstein	$O(2N)$	26.16	26.69
	✓	Average L2	$O(J^2N)$	26.11	26.64
	✓	Match prob	$O(J^2N)$	**26.28**	**26.77**

Table 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.

Method	HNM	Image-to-text R-P	Image-to-text R@1	Text-to-image R-P	Text-to-image R@1
VSE0	✗	22.4	44.2	22.6	32.7
PVSE K=1	✓	22.3	40.9	20.5	31.7
PVSE K=2	✓	19.7	47.3	21.2	28.0
PVSE K=4	✓	18.4	**47.8**	19.9	34.4
PCME μ only	✗	24.7	46.4	25.6	**35.5**
PCME	✗	**26.3**	46.9	**26.8**	35.2

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

PVSE with HNM tends to obtain better performances than VSE0 under the R@1 metric, with 47.8% for K=4, compared to 44.2% for VSE0. However, under the R-Precision metric, we observe all PVSE models with HNM are worse than VSE0 (R-Precision drops from 22.4% for VSE0 to 18.4% for PVSE K=4). It seems that PVSE with HNM tends to retrieve items based on diversity, rather than precision. We conjecture that the HNM is designed to optimize the R@1 performances; more details in §E.2. Comparing PVSE with different values of K, we note that increasing K does not always bring about performance gains under the R-Precision metric (20.5%, 21.2% and 19.9% for K=1,2,4, respectively, for t2i), while the improvement is more pronounced under the R@1 metric. Finally, PCME provides the best performances on both R-Precision and R@1 metrics, except for the R@1 score for i2t. PCME also improves upon its deterministic version, PCME μ-only, with some margin: +1.6 pp and +1.2 pp on i2t and t2i R-Precision scores, respectively.

4.3. Results on COCO

As we have identified potential problems with measuring performance on COCO (§4.1.2), we report the results with our Plausible-Match R-Precision (PMRP) metrics (§4.1.1) that captures the model performances more accurately than the widely-used R@k metrics. Table 3 shows the results

Method	1K Test Images				5K Test Images			
	i2t		t2i		i2t		t2i	
	PMRP	R@1	PMRP	R@1	PMRP	R@1	PMRP	R@1
VSE++ [9]	-	64.6	-	52.0	-	41.3	-	30.3
PVSE K=1 [45]	40.3*	66.7	41.8*	53.5	29.3*	41.7	30.1*	30.6
PVSE K=2 [45]	42.8*	69.2	43.6*	55.2	31.8*	45.2	32.0*	32.4
VSRN [25]	41.2*	76.2	42.4*	62.8	29.7*	53.0	29.9*	40.5
VSRN + AOQ [4]	44.7*	**77.5**	45.6*	**63.5**	33.0*	**55.1**	33.5*	**41.1**
PCMEμ only	**45.0**	68.0	45.9	54.6	34.0	43.5	34.3	31.7
PCME	**45.0**	68.8	**46.0**	54.6	**34.1**	44.2	**34.4**	31.9

Table 3. **Comparison on MS-COCO.** PMRP stands for the Plausible Match R-Precision and R@1 for Recall@1. "$*$" denotes results produced by the published models.

with state-of-the-art COCO retrieval methods. We observe that the stochastic version of PCME performs better than the deterministic variant (μ only) across the board. In terms of the R@1 metric, PVSE K=2 [45], VSRN [25] and AOQ [4] work better than PCME (*e.g.* 45.2%, 53.0%, 55.1% versus 44.2% for the 5K, i2t task). However, on the more accurate PMRP metric, PCME outperforms previous methods with some margin (*e.g.* 31.8%, 29.7%, 33.0% versus 34.1% for the 5K, i2t task). The results on two metrics imply that PCME retrieves the plausible matches much better than previous methods do. The full results can be found in §E.

4.4. Understanding the learned uncertainty

Having verified the retrieval performance of PCME, we now study the benefits of using probabilistic distributions for representing data. We show that the learned embeddings not only represent the inherent uncertainty of data but also enable set algebras among samples that roughly correspond to their semantic meanings.

Measuring uncertainty with σ. In an automated decision process, it benefits a lot to be able to represent uncertainty. For example, the algorithm may refrain from making a decision based on the uncertainty estimates. We show that the learned cross-modal embeddings capture the inherent uncertainty in the instance. We measure the instance-wise uncertainty for all query instances by taking the geometric mean over the $\sigma \in \mathbb{R}^D$ entries (§3.2.1). We then compute the average R@1 performances in each of the 10 uncertainty bins. Figure 6 plots the correlation between the uncertainty and R@1 on the COCO test set. We observe performance drops with increasing uncertainty. In §F.2, we visualize which word affects more to uncertainty. Example uncertain instances and their retrieval results are in §F.3.

2D visualization of PCME. To visually analyze the behavior of PCME, we conduct a 2D toy experiment by using 9 classes of the CUB Captions (details in §C.3). Figure 5 visualizes the learned image and caption embeddings. We also plot the embedding for the most generic caption for the CUB Captions dataset, "this bird has <unk> <unk> ...",

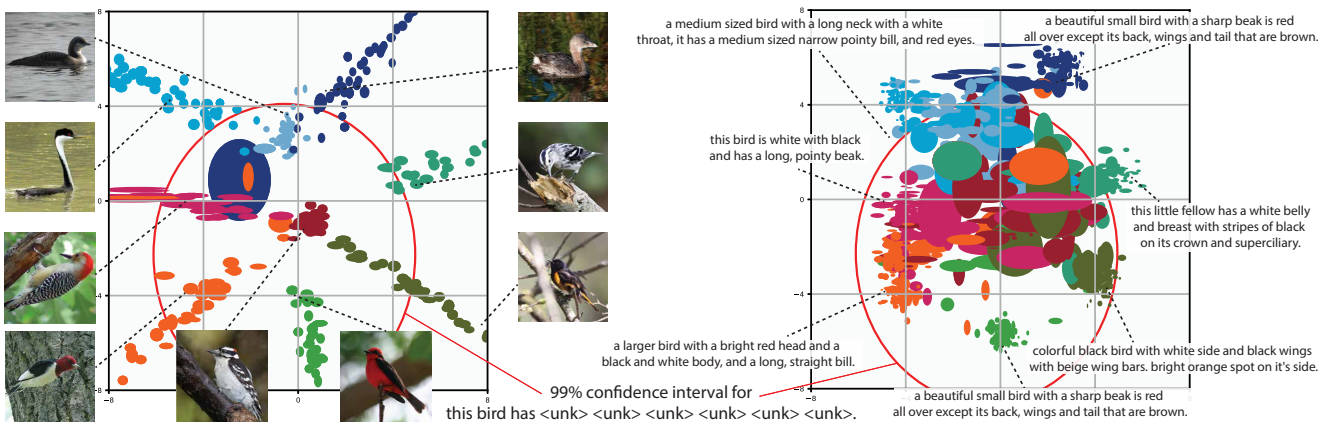


Figure 5. **Visualization of the probabilistic embedding.** The learned image (left) and caption (right) embeddings on 9 subclass of CUB Captions. Classes are color-coded. Each ellipse shows the 50% confidence region for each embedding. The red ellipse corresponds to the generic CUB caption, "this bird has <unk> $\cdots$ <unk>" with 99% confidence region.

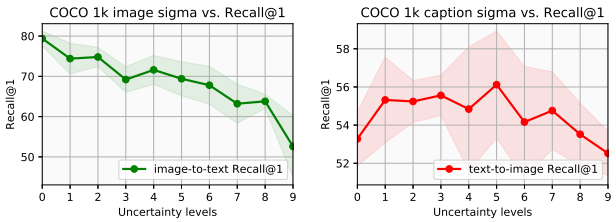


Figure 6. σ **versus performance.** Performance of PCME at different per-query uncertainty levels in COCO 1k test set.

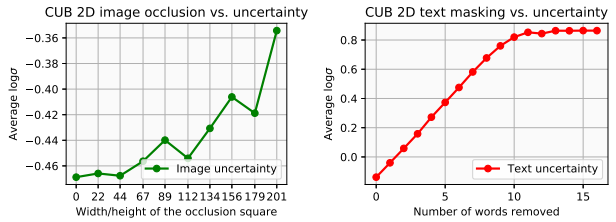


Figure 7. σ **captures ambiguity.** Average σ values at different ratios of erased pixels (for images) and words (for captions).

where <unk> is a special token denoting the absence of a word. This generic caption covers most of the caption variations in the embedding space (red ellipses).

Set algebras. To understand the relationship among distributions on the embedding space, we artificially introduce different types of uncertainties on the image data. In Figure 8, we start from two bird images and perform erasing and mixing transformations [59]. On the embedding space, we find that the mixing operation on the images results in embeddings that cover the *intersection* of the original embeddings. Occluding a small region in input images, on the other hand, amounts to slightly wider distributions, indicating an *inclusion* relationship. We quantitatively verify that the sigma values positively correlate with the ratio of erased pixels in Figure 7. In COCO, we observe a similar behavior (shown in §F.1). We discover another positive correlation

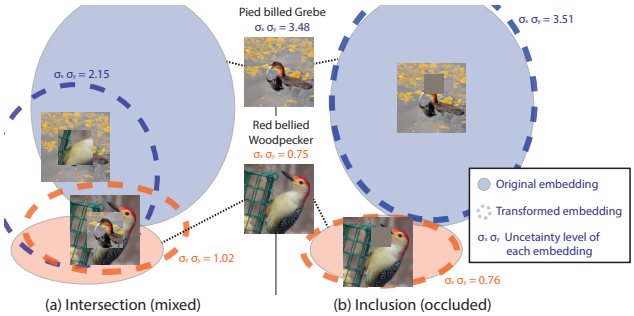


Figure 8. **Set algebras.** For two images, we visualize the embeddings for either erased or mixed samples. Mixing (left) and erasing (right) operations roughly translate to the intersection and inclusion relations between the corresponding embeddings.

between the caption ambiguity induced by erasing words and the embedding uncertainty.

5. Conclusion

We introduce Probabilistic Cross-Modal Embedding (PCME) that learns probabilistic representations of multi-modal data in the embedding space. The probabilistic framework provides a powerful tool to model the widespread one-to-many associations in image-caption pairs. To our knowledge, this is the first work that uses probabilistic embeddings for a multi-modal task. We extensively ablate our PCME and show that not only it improves the retrieval performance over its deterministic counterpart, but also provides uncertainty estimates that render the embeddings more interpretable.

Acknowledgements

We thank our NAVER AI Lab colleagues for valuable discussions. All experiments were conducted on NAVER Smart Machine Learning (NSML) [19] platform.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3, 4

[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. 5

[3] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proc. CVPR*, pages 5710–5719, 2020. 2

[4] Tianlang Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. In *Proc. ECCV*, 2020. 7

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 2, 6

[6] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 3

[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 5

[8] 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

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proc. BMVC*, 2018. 1, 2, 3, 6, 7

[10] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Proc. NeurIPS*, pages 2121–2129, 2013. 2, 3

[11] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *Proc. ECCV*, pages 529–545. Springer, 2014. 2

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proc. CVPR*, 2006. 3

[13] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 3, 5

[15] 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. 5

[16] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *Proc. CVPR*, pages 2310–2318, 2017. 2

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, pages 3128–3137, 2015. 6

[18] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Proc. NeurIPS*, pages 1889–1897, 2014. 2

[19] Hanjoo Kim, Minkyu Kim, Dongjoo Seo, Jinwoong Kim, Heungseok Park, Soeun Park, Hyunwoo Jo, KyungHyun Kim, Youngil Yang, Youngkwan Kim, et al. NSML: Meet the MLaaS platform with a real-world case study. *arXiv preprint arXiv:1810.09957*, 2018. 8

[20] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proc. ICLR*, 2014. 3

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Proc. ICLR*, 2014. 3

[22] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *Proc. ICLR*, 2017. 2

[23] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014. 2

[24] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proc. ECCV*, 2018. 2, 5

[25] 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. 1, 2, 7

[26] Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. Smoothing the geometry of probabilistic box embeddings. In *Proc. ICLR*, 2019. 2

[27] 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 *Proc. ECCV*, 2014. 6

[28] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proc. ACM-MM*, page 3–11, 2019. 2, 5

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *Proc. ICLR*, 2017. 5

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proc. NeurIPS*, pages 13–23, 2019. 2, 5

[31] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proc. CVPR*, pages 10437–10446, 2020. 2, 5

[32] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 2

[33] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Proc. ECCV*, 2020. 6

[34] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proc. CVPR*, pages 299–307, 2017. 2, 5

[35] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proc. EMNLP*, pages 1059–1069, 2014. 2

[36] Dat Quoc Nguyen, Ashutosh Modi, Stefan Thater, Manfred Pinkal, et al. A mixture model for learning multi-sense word embeddings. In *Proc. of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 121–127, 2017. 2

[37] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. In *Proc. ICLR*, 2019. 2, 3, 4

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proc. EMNLP*, 2014. 3, 5

[39] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proc. CVPR*, pages 49–58, 2016. 6

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. NeurIPS*, pages 91–99, 2015. 2

[41] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proc. CVPR*, pages 8247–8255, 2019. 2

[42] Tyler Scott, Karl Ridgeway, and Michael Mozer. Stochastic prototype embeddings. *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2019. 2

[43] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *ICCV*, 2019. 2

[44] Anna Silnova, Niko Brummer, Johan Rohdin, Themos Stafylakis, and Lukas Burget. Probabilistic embeddings for speaker diarization. In *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, pages 24–31, 2020. 2

[45] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proc. CVPR*, pages 1979–1988, 2019. 1, 2, 3, 4, 5, 6, 7

[46] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *Proc. ECCV*, 2020. 2

[47] Christopher Thomas and Adriana Kovashka. Preserving semantic neighborhoods for robust cross-modal retrieval. In *Proc. ECCV*, 2020. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, pages 5998–6008, 2017. 4

[49] Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating self-expression and visual content in hashtag supervision. In *Proc. CVPR*, pages 5919–5927, 2018. 2

[50] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In *Proc. ICLR*, 2015. 2

[51] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proc. CVPR*, pages 5005–5013, 2016. 1, 2

[52] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proc. ICML*, 2020. 4

[53] 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 *Proc. ICCV*, 2019. 2, 5

[54] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proc. CVPR*, 2020. 2, 5

[55] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 2, 6

[56] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *Proc. CVPR*, 2021. 2

[57] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proc. CVPR*, pages 4582–4591, 2017. 6

[58] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *ACL*, 2:67–78, 2014. 1

[59] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proc. ICCV*, 2019. 8

[60] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z. Li. Context-aware attention network for image-text retrieval. In *Proc. CVPR*, 2020. 2, 5