

# PROBABILISTIC MULTIMODAL REPRESENTATION LEARNING

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Learning multimodal representations is a requirement for many tasks such as image-caption retrieval. Previous work on this problem has only focused on finding good vector representations without any explicit measure of uncertainty. In this work, we argue and demonstrate that learning multimodal representations as probability distributions can lead to better representations, as well as providing other benefits such as adding a measure of uncertainty to the learned representations. We show that this measure of uncertainty can capture how confident our model is about the representations in the multimodal domain, i.e, how clear it is for the model to retrieve/predict the matching pair. We experiment with similarity metrics that have not been traditionally used for the multimodal retrieval task, and show that the choice of the similarity metric affects the quality of the learned representations.

## 1 INTRODUCTION

Representation learning is a core problem of machine learning that deals with the embedding of raw input data (e.g., an image or a document) into a latent space. Learning such embeddings is a pre-requisite for automated information processing tasks, such as language modeling (Mikolov et al., 2013), face recognition (Schroff et al., 2015), and image retrieval (Ustinova & Lempitsky, 2016). In recent years, there has been a surge of interest in deep representation learning, with remarkable empirical successes in a variety of applications. Many have gone beyond representing data of a single type, and have proposed methods to learn joint embeddings that simultaneously encode several different modalities (such as image, video, text, speech, or audio) into a common space (Kiros et al., 2014; Vendrov et al., 2016; Faghri et al., 2018; Harwath et al., 2018; Lee et al., 2018; Zablocki et al., 2018; Lu et al., 2019; Liu et al., 2019). Multimodal techniques can learn richer representations that enable comparison across modalities, which is required by applications such as image captioning (Socher et al., 2014; Karpathy & Fei-Fei, 2015) and language-based image retrieval (Faghri et al., 2018; Lee et al., 2018).

Most representation learning techniques learn the latent embeddings by mapping each data item (e.g., a single image-caption pair) to a dense vector representing a point in a  $k$ -dimensional space, with no explicit encoding of uncertainty. In the case of word embeddings, a clear drawback is that such point estimates cannot easily address issues such as polysemy, i.e., when a word such as *star* can have two or more different meanings, i.e., *movie star* and *celestial body*. A few recent studies have advocated for density-based or probabilistic word embeddings that capture such nuances of meaning (Vilnis & McCallum, 2014; Athiwaratkun & Wilson, 2017; 2018; Athiwaratkun et al., 2018). In a multimodal setting, Mukherjee & Hospedales (2016) show that drawing on density-based word and image representations improves zero-shot image classification.

Inspired by findings of density-based learning approaches, we ask the question whether joint multimodal representations benefit from such methods in a more general setting. Specifically, we seek to understand if we can learn higher-quality multimodal representations based on probabilistic embeddings of captions and images, as opposed to the standard joint spaces learned over vector-based representations. We propose to learn Gaussian embeddings of captions and images, by using an extension of the word learning model of Vilnis & McCallum (2014). From these probabilistic embeddings, we learn a common latent space that can capture nuances of meaning with respect to image-caption correspondences. We evaluate our models by measuring their performance on the

standard task of multimodal (image–text or text–image) retrieval, and compare them with a baseline (vector-based) model. Our results confirm the superiority of the probabilistic representations, both in terms of overall retrieval performance, and in terms of capturing nuances of meaning when it comes to choosing the most appropriate response (e.g., a caption) for a given query (e.g., an image).

Our contributions are as follows: To the best of our knowledge, we are the first to propose models for learning joint multimodal representations from Gaussian embeddings of images and captions. Through extensive experimentation, we show that non-standard similarity metrics (that have not been previously used in deep representation learning) can generate high-quality representations. Additionally, we provide an explicit measure of uncertainty (or entropy), and show that the measure indeed correlates with the model’s confidence in retrieving the correct match for a given query. Finally, we provide a qualitative analysis and show that with probabilistic representations the retrieved images/captions are more aligned with the query in terms of entropy.

## 2 MODEL

In this work, we propose models that learn to represent images and textual captions as probability distributions. These learned distributions are fused to map images and captions into a joint latent space, enriching the representations and enabling cross-modality comparisons. To train our model we assume we have pairs of images and corresponding captions,  $(i_n, c_n)$  for  $n \in \{1, \dots, N\}$  where  $N$  is the total number of pairs in the dataset. A pair  $(i_j, c_k)$  forms a *matching* pair if  $j = k$  and a *non-matching* pair if  $j \neq k$ . For ease of exposition, we use  $(i, c)$  to refer to a matching pair and  $(i', c)$  and  $(i, c')$  to represent a non-matching pair. We use contrastive learning 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. Specifically, we use the hinge-based triplet ranking loss (Chechik et al., 2010) with semi-hard negative mining (Schroff et al., 2015), defined for a positive pair as:

$$\mathcal{L}(i, c) = \max_{c'} [\alpha + \text{sim}(i, c') - \text{sim}(i, c)]_+ + \max_{i'} [\alpha + \text{sim}(i', c) - \text{sim}(i, c)]_+, \quad (1)$$

where  $[x]_+ \triangleq \max(x, 0)$ ,  $\alpha$  is a hyperparameter for the margin, and  $\text{sim}(i, c)$  is a measure of similarity. The total loss is the sum of  $\mathcal{L}(i, c)$  for all positive pairs, i.e.:

$$\mathcal{L} = \sum_{(i, c)} \mathcal{L}(i, c). \quad (2)$$

We build upon the multimodal representation learning system of Faghri et al. (2018) as it has a simple architecture that does not rely on object detectors for capturing the image encodings and it has good performance given its simplicity. Similar to the VSE++ model (Faghri et al., 2018), our model consists of a pretrained image encoder and a GRU-based caption encoder with learnable parameters  $\theta_\phi$  and  $\theta_\psi$ , respectively. We represent the encoding for image  $i$  as  $\phi(i; \theta_\phi) \in \mathbb{R}^{D_\phi}$  and the encoding for caption  $c$  as  $\psi(c; \theta_\psi) \in \mathbb{R}^{D_\psi}$ . We use these encodings to learn probabilistic representations for the images and captions in a  $D$  dimensional space. Following Vilnis & McCallum (2014), we assume diagonal multivariate ellipsoidal or spherical Gaussian distributions to model our probabilistic representations. We train three variations of our model: One with Gaussian caption embeddings and vector image representations (GCE; Section 2.1), one with Gaussian image embeddings and vector caption representations (GIE; Section 2.2), and one with Gaussian caption and image embeddings (GCIE; Section 2.3). Throughout the paper, we use  $\mathcal{N}(\mu, \Sigma)$  to indicate a Gaussian distribution with mean  $\mu$  and covariance matrix  $\Sigma$ . We use  $\text{diag}(\mathbf{a})$  to refer to a diagonal matrix, where  $\mathbf{a}$  is a vector containing the elements on the diagonal.

### 2.1 GAUSSIAN CAPTION EMBEDDING (GCE)

In this model, images are represented as vectors  $i$ , and captions as Gaussian distributions  $\mathcal{N}(\mu_c, \Sigma_c)$  with a per-caption mean  $\mu_c$  and a per-caption covariance matrix  $\Sigma_c$ , where  $\Sigma_c = \text{diag}(\sigma_c^2)$ . Image representations are learned via a linear transformation (fully-connected layer) of the image encoding  $\phi(i; \theta_\phi)$ . To learn the sufficient statistics of the Gaussian caption distributions, we use two caption

encoders with no shared parameters, and add fully connected layers to transform the encodings:

$$\mathbf{i} = f(i; W_f, \boldsymbol{\theta}_\phi) = W_f \phi(i; \boldsymbol{\theta}_\phi), \quad (3)$$

$$\boldsymbol{\mu}_c = g_\mu(c; W_{g_\mu}, \boldsymbol{\theta}_{\psi_\mu}) = W_{g_\mu} \psi_\mu(c; \boldsymbol{\theta}_{\psi_\mu}), \quad (4)$$

$$\boldsymbol{\sigma}_c^2 = g_{\sigma^2}(c; W_{g_{\sigma^2}}, \boldsymbol{\theta}_{\psi_{\sigma^2}}) = W_{g_{\sigma^2}} \psi_{\sigma^2}(c; \boldsymbol{\theta}_{\psi_{\sigma^2}}), \quad (5)$$

where  $D_{\psi_\mu} = D_{\psi_{\sigma^2}} = D_\psi$ ,  $W_f \in \mathbb{R}^{D_\phi \times D}$ , and  $W_{g_\mu}, W_{g_{\sigma^2}} \in \mathbb{R}^{D_\psi \times D}$ . With no fine-tuning of the image encoder, the model parameters are:  $\boldsymbol{\theta} \triangleq [\boldsymbol{\theta}_{\psi_\mu}; \boldsymbol{\theta}_{\psi_{\sigma^2}}; W_f; W_{g_\mu}; W_{g_{\sigma^2}}]$ . When we fine-tune the image encoder,  $\boldsymbol{\theta}_\phi$  is added to the model parameter vector  $\boldsymbol{\theta}$ . The model parameters are learned by minimizing the loss function in equation 2 with the negative of Mahalanobis distance as the measure of similarity:<sup>1</sup>

$$\text{sim}(i, c) = -\sqrt{(\mathbf{i} - \boldsymbol{\mu}_c)^T \Sigma_c^{-1} (\mathbf{i} - \boldsymbol{\mu}_c)}. \quad (6)$$

## 2.2 GAUSSIAN IMAGE EMBEDDING (GIE)

In this model, captions are represented as vectors  $\mathbf{c}$ , and images as Gaussian distributions  $\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$  with a per-image mean  $\boldsymbol{\mu}_i$  and a per-image covariance matrix  $\Sigma_i$ , where  $\Sigma_i = \text{diag}(\boldsymbol{\sigma}_i^2)$ . Caption representations are learned via a linear transformation of the caption encoding  $\psi(c; \boldsymbol{\theta}_\psi)$ . To learn the sufficient statistics of the Gaussian image distributions, we add two fully connected layers (with no shared parameters) to transform the image encoding  $\phi(i; \boldsymbol{\theta}_\phi)$ . Note that unlike for the GCE model, here we use a single image encoding, but use two independent linear transformations to learn the parameters of the Gaussians:<sup>2</sup>

$$\boldsymbol{\mu}_i = f_\mu(i; W_{f_\mu}, \boldsymbol{\theta}_\phi) = W_{f_\mu} \phi(i; \boldsymbol{\theta}_\phi), \quad (7)$$

$$\boldsymbol{\sigma}_i^2 = f_{\sigma^2}(i; W_{f_{\sigma^2}}, \boldsymbol{\theta}_\phi) = W_{f_{\sigma^2}} \phi(i; \boldsymbol{\theta}_\phi), \quad (8)$$

$$\mathbf{c} = g(c; W_g, \boldsymbol{\theta}_\psi) = W_g \psi(c; \boldsymbol{\theta}_\psi), \quad (9)$$

where  $W_{f_\mu}, f_{\sigma^2} \in \mathbb{R}^{D_\phi \times D}$ , and  $W_g \in \mathbb{R}^{D_\psi \times D}$ . Without fine-tuning the image encoder, the model parameters are:  $\boldsymbol{\theta} = [\boldsymbol{\theta}_\psi; W_{f_\mu}; f_{\sigma^2}; W_g]$ , and when we fine-tune the image encoder,  $\boldsymbol{\theta}_\phi$  is added to the model parameter vector  $\boldsymbol{\theta}$ . The model parameters are learned by minimizing the loss function in equation 2 with the negative of Mahalanobis distance as the measure of similarity:

$$\text{sim}(i, c) = -\sqrt{(\mathbf{c} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{c} - \boldsymbol{\mu}_i)}. \quad (10)$$

## 2.3 GAUSSIAN CAPTION AND IMAGE EMBEDDING (GCIE)

In this model, both images and captions are represented as Gaussian distributions,  $\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$  and  $\mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c)$ , where  $\Sigma_i = \text{diag}(\boldsymbol{\sigma}_i^2)$  and  $\Sigma_c = \text{diag}(\boldsymbol{\sigma}_c^2)$ . We follow the same approach described in Sections 2.1 and 2.2 to learn the sufficient statistics for the image and caption representations, respectively:

$$\boldsymbol{\mu}_i = f_\mu(i; W_{f_\mu}, \boldsymbol{\theta}_\phi) = W_{f_\mu} \phi(i; \boldsymbol{\theta}_\phi), \quad (11)$$

$$\boldsymbol{\sigma}_i^2 = f_{\sigma^2}(i; W_{f_{\sigma^2}}, \boldsymbol{\theta}_\phi) = W_{f_{\sigma^2}} \phi(i; \boldsymbol{\theta}_\phi), \quad (12)$$

$$\boldsymbol{\mu}_c = g_\mu(c; W_{g_\mu}, \boldsymbol{\theta}_{\psi_\mu}) = W_{g_\mu} \psi_\mu(c; \boldsymbol{\theta}_{\psi_\mu}), \quad (13)$$

$$\boldsymbol{\sigma}_c^2 = g_{\sigma^2}(c; W_{g_{\sigma^2}}, \boldsymbol{\theta}_{\psi_{\sigma^2}}) = W_{g_{\sigma^2}} \psi_{\sigma^2}(c; \boldsymbol{\theta}_{\psi_{\sigma^2}}), \quad (14)$$

where  $D_{\psi_\mu} = D_{\psi_{\sigma^2}} = D_\psi$ ,  $W_{f_\mu}, W_{f_{\sigma^2}} \in \mathbb{R}^{D_\phi \times D}$ , and  $W_{g_\mu}, W_{g_{\sigma^2}} \in \mathbb{R}^{D_\psi \times D}$ . Without fine-tuning the image encoder, for the model parameters we have  $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\psi_\mu}; \boldsymbol{\theta}_{\psi_{\sigma^2}}; W_{f_\mu}; W_{f_{\sigma^2}}; W_{g_\mu}; W_{g_{\sigma^2}}]$ , and when we fine-tune the image encoder,  $\boldsymbol{\theta}_\phi$  is added to the model parameter vector  $\boldsymbol{\theta}$ . We learn the model parameters by minimizing the loss function in equation 2 with one of the following three similarity functions to estimate  $\text{sim}(i, c)$ :

<sup>1</sup>The leading negative sign is to turn the distance measure into a measure of similarity. We use Mahalanobis distance because we need to measure similarity between a point and a probability distribution.

<sup>2</sup>Unlike the caption encoder parameters that are learned from scratch, for the image encoder we start from a pretrained network. As such, there is no need to use two such networks (especially with no fine-tuning). With fine-tuning, using two separate networks as image encoders would require a significant increase in the number of model parameters.

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

$$-\frac{1}{2} \left( \text{tr}(\Sigma_c^{-1} \Sigma_i) + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_c)^T \Sigma_c^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_c) - D + \ln \left( \frac{\det \Sigma_c}{\det \Sigma_i} \right) \right), \quad (15)$$

where  $\text{tr}(\mathbf{A})$  and  $\det(\mathbf{A})$  are the trace and determinant of matrix  $\mathbf{A}$ , respectively.

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

$$-\min(\text{KL}(\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i) || \mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c)), \text{KL}(\mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c) || \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i))), \quad (16)$$

where  $\text{KL}(P || Q)$  is the KL divergence from probability distribution  $Q$  to  $P$ .

3. **Negative 2-Wasserstein distance.** Wasserstein distance is a symmetric distance metric between probability distributions. We use the 2nd order Wasserstein (2-Wasserstein) because it can be evaluated using the following closed form for Gaussians (Mallasto & Feragen, 2017):

$$-\sqrt{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_c\|^2 + \text{tr} \left( \Sigma_i + \Sigma_c - 2 \left( \Sigma_i^{\frac{1}{2}} \Sigma_c \Sigma_i^{\frac{1}{2}} \right)^{\frac{1}{2}} \right)}. \quad (17)$$

### 3 EVALUATION

#### 3.1 EXPERIMENTAL SETUP

**Data.** For training and evaluation, we use the (Microsoft) MS-COCO dataset (Lin et al., 2014). We work with the splits from Karpathy & Fei-Fei (2015), containing 82,783 training, 5000 validation, and 5000 test images, where each image is associated with five manually-provided captions that describe the contents of the image. Following Faghri et al. (2018), we add to our training data an additional set of 30,504 images (and their captions) that were in the original validation set of MS-COCO, but have been left out from the Karpathy & Fei-Fei (2015) split.

**Training details.** We use the VSE++ (Faghri et al., 2018) code<sup>3</sup> as our baseline with the modifications described in Section 2. Specifically, for learning probabilistic image representations in GIE and GCIE, we replace the last logit layer of the image encoder and add two fully connected layers to learn the means and the log variances<sup>4</sup> of the Gaussian image representations. For learning the probabilistic caption representations for GCE and GCIE, we duplicate the GRU-based encoder and use one encoder for estimating the means and the other one for learning the log variances of the caption representation. For spherical distributions, we take the mean of the log variances over the embedding dimensions. Following Vilnis & McCallum (2014) we bound all variance values to the range of  $[0.1, 10]$ .

We train and test our models following the same settings as in VSE++, i.e., we use pretrained ResNet 152 (He et al., 2016) with  $D_\phi = 2048$  as our base image encoder and a GRU-based caption encoder with  $D_\psi = 1024$  an input size of 300, and train our models for 15 epochs with a learning rate of  $2e-4$ , followed by another 15 epochs with a lower learning rate of  $2e-5$ . The dimensionality of our joint embedding space is  $D = 1024$ . When we fine-tune the image encoder in our model, we start from the trained models with a fixed image encoder and train for 15 epochs at the learning rate of  $2e-5$  for all models, except GIE spherical, where we use the learning rate of  $2e-6$ . We use the hinge-based triplet ranking loss (equation 1) with  $\alpha = 0.2$ , with the different similarity metrics described in Sections 2.1–2.3.

<sup>3</sup><https://github.com/fartashf/vsepp>

<sup>4</sup>We estimate the log variances instead of variances to ensure numerical stability.



Table 1: Recall values of all models (no fine-tuning) on MS-COCO test set; results are averages over five folds of 1000 images each; best performances are shown in **boldface**; performances similar to or better than the baseline VSE++ are shown in *italics*.

Model	similarity metric	image to text			text to image		
		R@1	R@5	R@10	R@1	R@5	R@10
VSE++	cosine	58.3	86.1	93.3	43.6	77.6	87.8
GCE spherical	Mahalanobis	58.1	86.1	<b>93.8</b>	43.3	78.0	88.3
GCE ellipsoidal	Mahalanobis	57.3	85.5	93.1	43.9	77.9	88.4
GIE spherical	Mahalanobis	58.0	85.8	92.7	43.6	77.6	88.1
GIE ellipsoidal	Mahalanobis	58.2	86.0	93.5	44.2	78.2	88.5
GCIE spherical	KL	56.4	85.0	92.4	42.0	76.4	87.3
GCIE ellipsoidal	KL	57.8	86.6	94.0	44.8	79.0	89.0
GCIE spherical	min KL	55.9	84.2	92.6	41.8	76.3	87.2
GCIE ellipsoidal	min KL	58.2	85.6	93.1	44.9	78.6	89.0
GCIE spherical	Wasserstein	57.3	85.9	93.2	43.8	77.7	88.1
GCIE ellipsoidal	Wasserstein	<b>59.0</b>	<b>87.0</b>	93.6	<b>45.1</b>	<b>79.2</b>	<b>89.2</b>

**Evaluation.** As is standard in multimodal representation learning, we report  $R(\text{ecall})@K$  (with  $K = 1, 5, 10$ ) for both image-to-text and text-to-image retrieval tasks. All reported results are averages over five folds of 1000 test images each.

To measure the entropy of our representations in the joint embedding space, we use  $\log(\det \Sigma_i)$  and  $\log(\det \Sigma_c)$  for images and captions, respectively, since the differential entropy of a multivariate Gaussian with covariance matrix  $\Sigma \in \mathbb{R}^{d \times d}$  can be calculated as:

$$\frac{1}{2} (d + d \log(2\pi) + \log(\det \Sigma)). \quad (18)$$

This measure can help us understand the uncertainty associated with retrieving an image given a query caption, or a caption given a query image.

### 3.2 QUANTITATIVE RESULTS

Table 1 reports retrieval results for the original point-based vector representations learned by VSE++, as well as variations of the density-based representations learned by our models. Results are reported for test set, with the best configurations of each model identified by looking at the sum of recall values (for text-to-image and image-to-text) for the validation set. As can be seen, the best performances (given in **boldface**) are generally achieved by our ellipsoidal GCIE model, using the Wasserstein distance. These results especially point to the Wasserstein distance as an appropriate metric that has been largely overlooked in representation learning. We can see that in many cases, our models outperform VSE++ (*italicized* recall values), especially for text-to-image retrieval.

Table 2 shows the retrieval results, with fine-tuning of the image encoding network. Here again we can see that several of our models outperform VSE++. In particular, the GCIE ellipsoidal model with the min KL and Wasserstein distances produce the best results. Once again, these results provide evidence for the effectiveness of probabilistic representations in the multimodal setting, and further suggest the importance of considering new metrics for representation learning.

In both Tables 1 and 2, we observe that the GCIE ellipsoidal models (with different similarity metrics) all have notably higher recall values compared to their spherical counterparts (on both image-to-text and text-to-image retrieval tasks). We believe this is due to the fact that with ellipsoidal image distributions, the model can better capture the variations of variances along different dimensions and for different images.

### 3.3 QUALITATIVE RESULTS

In Figure 1 we provide four examples where we take crops from random images in the test set. We truncate the captions for these images to match the crops, and we feed them one pair at a time to both the GCIE and the VSE++ models to compare the rankings given by the two models. Since GCIE

Table 2: Recall values of all models (with fine-tuning) on MS-COCO test set; results are averages over five folds of 1000 images each; best performances are shown in **boldface**; performances similar to or better than the baseline VSE++ are shown in *italics*.

Model	similarity metric	image to text			text to image		
		R@1	R@5	R@10	R@1	R@5	R@10
VSE++	cosine	64.6	90.0	95.7	52.0	84.3	92.0
GCE spherical	Mahalanobis	63.9	89.4	95.4	51.6	84.1	92.2
GCE ellipsoidal	Mahalanobis	63.4	89.0	95.5	52.2	84.2	92.4
GIE spherical	Mahalanobis	54.4	82.7	91.3	40.0	73.7	85.0
GIE ellipsoidal	Mahalanobis	61.8	88.3	95.0	48.7	82.2	91.1
GCIE spherical	KL	61.0	88.9	95.0	49.9	83.3	91.7
GCIE ellipsoidal	KL	64.7	90.8	96.3	51.6	84.0	92.2
GCIE spherical	min KL	60.4	87.7	94.6	49.0	82.7	91.5
GCIE ellipsoidal	min KL	66.4	<b>91.3</b>	<b>96.5</b>	<b>52.5</b>	<b>84.6</b>	<b>92.7</b>
GCIE spherical	Wasserstein	61.8	88.8	94.7	50.9	83.7	92.1
GCIE ellipsoidal	Wasserstein	<b>66.9</b>	90.6	96.2	52.4	<b>84.6</b>	<b>92.7</b>

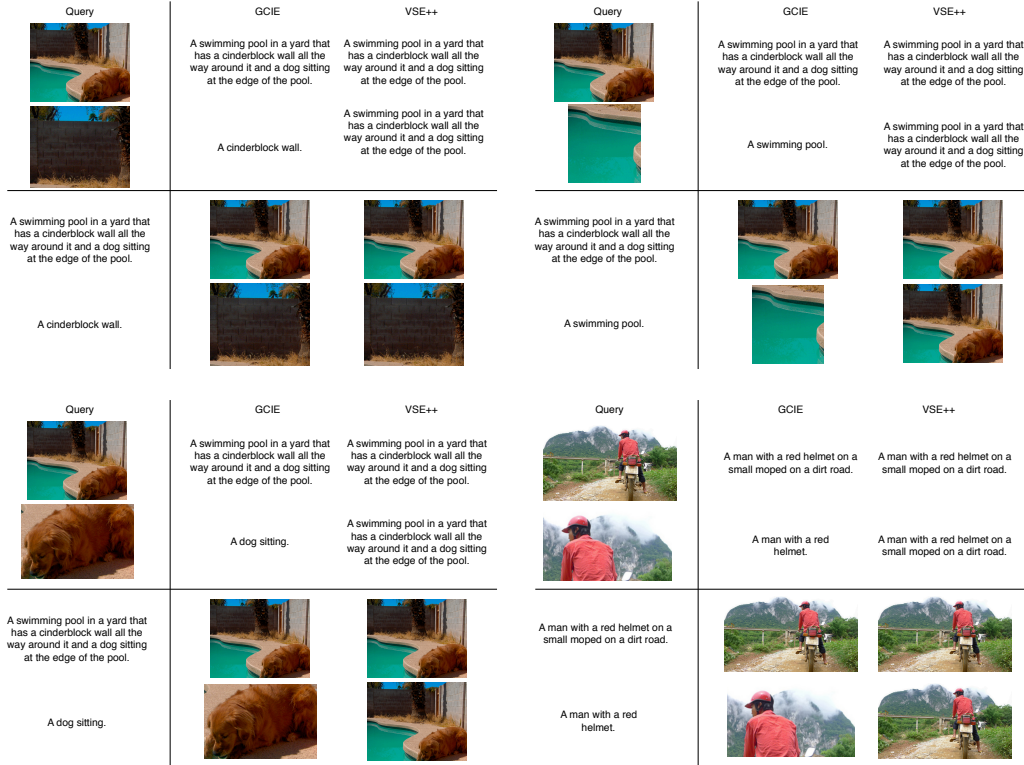




Figure 1: Retrieval results from our model and VSE++ on image crops and associated caption truncations on random images from the test set.

Table 3: Image/caption pairs with their crops/truncations and their associated measure of entropy.

caption	image	measure of entropy	
		caption	image
A man with a red helmet on a small moped on a dirt road.		18.19	26.14
A man with a red helmet.		15.51	19.05
A swimming pool in a yard that has a cinderblock wall all the way around it and a dog sitting at the edge of the pool.		19.20	21.51
A swimming pool.		10.52	12.39
A cinderblock wall.		15.48	14.66
A dog sitting.		14.28	20.65

can capture the level of details through the covariance matrices of the Gaussian representations, we expect its retrieval results to be better than VSE++. The results confirm that this is indeed the case. For instance, in response to an image depicting a pool, a brickwall, a tree, and a dog, both GCIE and VSE++ provide the long detailed caption as the match. But, with the cropped image that depicts the brickwall only, GCIE (but not VSE++) is returning the short caption that is a more appropriate match. The same pattern is observed for all four sets of examples in the figure.

Recall that our uncertainty measure in equation 18 estimates the uncertainty or entropy associated with retrieving an image given a query caption (or vice versa). To better depict the relationship between the image crops and the caption truncations in Figure 1, we provide the measure of entropy for them in Table 3. Interestingly, this measure can capture how uncertain the GCIE model is for retrieving the matching caption or image for a given query. Specifically, we can see for a cropped image when we have a more focused subject in the image, the entropy is lower. Similarly for captions, when we have a truncation the model is more certain in grounding the shorter caption in the image and retrieving the matching image.

### 3.4 ANALYSIS OF LEARNED PROBABILISTIC REPRESENTATIONS

In our qualitative analysis in Section 3.3, we observe that our probabilistic representations are better at capturing nuances of meaning, compared to those learned by a model such as VSE++. In this section, we perform similar experiments using more realistic data on a larger scale, and report performance of our best model (GCIE ellipsoidal with Wasserstein as similarity metric) vs. VSE++.

In popular image retrieval datasets, captions vary in quality and degree of specificity. Often times captions describe only a particular (salient) part of an image, and leave other parts of the image unmentioned. We believe a probabilistic representation can better model the level of specificity in the visual and linguistic aspects of the data. We demonstrate this by testing our model’s ability in differentiating between parts of an image versus the whole, given a more or less specific query. To this end, we use the Visual Genome dataset (Krishna et al., 2016), which comes with annotated pairs of image regions and descriptions (mostly phrases). We remove the portion of Visual Genomen images that overlap with those in MS-COCO to avoid any unwanted information leak (since our model has been trained on MS-COCO). We randomly pick a set of images from the remaining Visual Genome data, and for each image select  $M = 3$  largest crops that cover an area no bigger than 20% of the image — we set this area constraint so that there are meaningful differences between the crop and the image. We try to choose crops with minimal overlap, but still observe some overlap between the selected crops. Note that the image regions (crops) come with their own linguistic annotation/description (usually in the form of a phrase), but the whole image does not have an associated caption in Visual Genome. As such, we form a description for each of our selected images by concatenating the phrases of its three selected crops.

We test our model in two conditions: (a) image retrieval, where we ask both our model (GCIE-best) and VSE++ to retrieve one of two images (a full image and one of its selected crops) in response to a text query that either describes the full image or the selected crop; (b) caption retrieval, where we ask the two models to retrieve one of two captions in response to an image. For image retrieval, out of a sample of 2,850 test items, when given the crop description as query, GCIE-best retrieves the correct response 61% of the time, vs. 58% accuracy for VSE++. When given the full-image description as query, GCIE-best correctly retrieves the original image at an accuracy of 77%, while VSE++ is correct only 69% of the time. When performing caption retrieval for the same 2,850 test items, VSE++ shows a strong bias towards shorter captions, correctly retrieving the crop description (a short phrase) when given the crop as query 97% of the time, compared to 82% accuracy for our model. But, when the full images are given as query, VSE++ has a very low accuracy of 13% while our model returns the correct description 53% of the time. These results further suggest that a probabilistic representation is better suited for capturing nuances of meaning.

## 4 CONCLUSIONS

In this paper, we proposed new models that learn image and text representations as probability distributions. We showed that by fusing such density-based representations we can learn higher-quality joint embeddings that can result in better multimodal retrieval performance. Through extensive evaluation of our models, we observed that the best model is one where both images and captions are represented as probability distributions, and that symmetric similarity metrics with bounded values are the most effective. Additionally, our qualitative analysis reveals an interesting behaviour of such a probabilistic model: Unlike a vector-based model, ours is sensitive to nuances of meaning, and is better at finding the most appropriate match (e.g., an image) with the right level of detail, given a query (e.g., a caption).

## REFERENCES

- Ben Athiwaratkun and Andrew Gordon Wilson. Multimodal Word Distributions. In *Annual Meeting of the Association for Computational Linguistics*, 2017.
- Ben Athiwaratkun and Andrew Gordon Wilson. Hierarchical Density Order Embeddings. In *International Conference on Learning Representations*, 2018.

- Ben Athiwaratkun, Andrew Gordon Wilson, and Anima Anandkumar. Probabilistic FastText for Multi-Sense Word Embeddings. In *Annual Meeting of the Association for Computational Linguistics*, 2018.
- 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.
- 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.
- David Harwath, Adria Recasens, Didac Suris, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *European Conference on Computer Vision*, 2018.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
- A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.
- Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models, 2014.
- 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. 2016. URL <https://arxiv.org/abs/1602.07332>.
- 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*, pp. 201–216, 2018.
- 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 (eds.), *European Conference on Computer Vision*, pp. 740–755. Springer International Publishing, 2014.
- Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *arXiv preprint arxiv:1907.13487*, 2019.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visual-linguistic representations for vision-and-language tasks. In *Conference on Neural Information Processing Systems*, 2019.
- Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes. In *Conference on Neural Information Processing Systems*, pp. 5660–5670, 2017.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Conference on Neural Information Processing Systems*, 2013.
- Tanmoy Mukherjee and Timothy Hospedales. Gaussian visual-linguistic embedding for zero-shot recognition. In *Empirical Methods in Natural Language Processing*, pp. 912–918, 2016.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.
- Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2, 2014.
- Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. In *Conference on Neural Information Processing Systems*, 2016.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-Embeddings of Images and Language. In *ICLR*, 2016.

Luke Vilnis and Andrew McCallum. Word Representations via Gaussian Embedding. In *ICLR*, 2014.

Éloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. Learning Multi-Modal Word Representation Grounded in Visual Context. In *Association for the Advancement of Artificial Intelligence*, 2018.