CALIBRATING PROBABILISTIC EMBEDDINGS FOR CROSS-MODAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

The core of cross-modal retrieval is to measure the content similarity between data of different modalities. The main challenge focuses on learning a shared representation space for multiple modalities where the similarity measurement can reflect the semantic closeness. The multiplicity of correspondences further escalates the challenge since all the possible matches should be ranked ahead of the negatives. Probabilistic embeddings are proposed to handle the multiplicity while suffering from similarity miscalibration. To address it, we propose to calibrate the similarity for probabilistic embeddings. The key idea is to estimate the density ratio between the distributions of the two modalities, and use it to calibrate the similarity measurement in the embedding space. To the best of our knowledge, we are the first to study the miscalibration in probabilistic embeddings. In addition, we further evaluate three pre-training tasks of language models, which is important for cross-modal but seldom investigated in previous studies. Extensive experiments as well as ablation studies on two benchmarks demonstrate its superior performance in tackling the multiplicity of cross-modal retrieval.

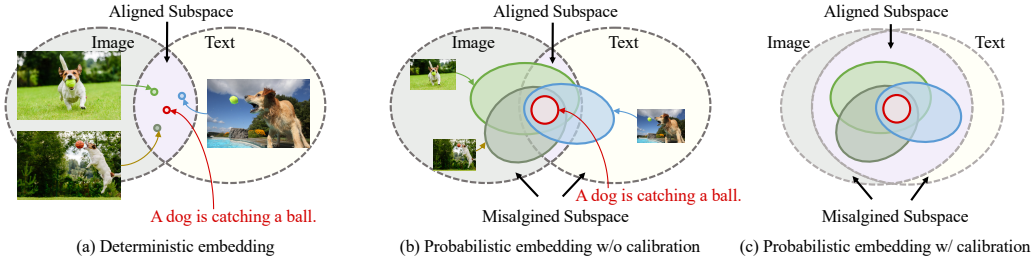
1 INTRODUCTION

Visual media and natural language are the two most prevalent modalities exhibiting information in our daily life. It is essential for computers to understand, match, and transform such cross-modal data. Cross-modal retrieval, a fundamental and crucial problem in multimodal learning, has attracted extensive attention in recent years. Typically, cross-modal retrieval requires a shared representation space that allows computing a similarity measurement between the query and the retrieved data.

Building a shared representation space for multiple modalities is challenging due to the heterogeneity across different modalities. The multiplicity of correspondences further escalates the challenge: an image potentially matches with a number of different texts. Conversely, given a caption, there may be multiple manifestations of the caption in visual forms (Fig. 1 (a)). The multiplicity poses new challenges for similarity measurement in cross modal retrieval since all the possible matches should be ranked ahead of the negatives. Most methods (Rasiwasia et al., 2010; Yan & Mikolajczyk, 2015; Wang et al., 2019; Wei et al., 2020) ignore the multiplicity and use deterministic functions to map a sample as a single point in the embedding space (Fig. 1 (a)). However, the single point representation can hardly handle the multiplicity. Recently, probabilistic embeddings (Oh et al., 2018; Chun et al., 2021) are proposed to map a sample as a Gaussian distribution in the embedding space, in order to cover a large area and increase the possibility to search for more possible matches.

The underlying assumption of probabilistic embeddings is that the two modalities are perfectly aligned and embeddings lie in a shared representation space. However, the assumption does not always hold in practice due to the modality heterogeneity (Rasiwasia et al., 2010), As a result, the embedding spaces are partially rather than fully aligned (Fig. 1 (b)). Besides, modality embeddings may easily fall out of the shared space when the variance is large and the overlap is small, yielding inaccurate similarity measurement between image-text pairs.

To narrow the aforementioned gap, we propose a novel calibration method for probabilistic embeddings in cross-modal retrieval. We propose to estimate the density ratio between the two distributions, and use it to calibrate the similarity measurement in the embedding space. The similarity calibration can effectively align the distributions between the two modalities and enlarge the shared representation space (Fig. 1 (c)), facilitating more accurate one-to-many matching. Our method can be seamlessly



(a) Deterministic embedding (b) Probabilistic embedding w/o calibration (c) Probabilistic embedding w/ calibration

Figure 1: (a) Deterministic embedding can hardly handle the multiplicity; b) Probabilistic embedding suffers from misaligned subspace; c) We propose a new method to enlarge the aligned subspace.

integrated into current contrastive learning frameworks where matched pairs will be pulled closer and unmatched pairs will be pushed away in the embedding space. In addition to the similarity calibration, we further evaluate three pre-training tasks of language models, which plays the crucial role for cross-modal retrieval but seldom investigated in previous studies. To summarize, our contribution is multi-fold: 1) To the best of our knowledge, we are the first to study the miscalibration issue in probabilistic embeddings. 2) We propose to use density ratio between the two modalities to calibrate the similarity measurement between image-text pairs, enlarging the shared representation space and facilitating more accurate one-to-many matching. 3) We investigate different pre-training tasks for cross-modal retrieval. 4) Extensive experiments as well as ablation studies on two benchmarks demonstrate its superior performance in tackling the multiplicity for cross-modal retrieval.

2 Related Work

Cross-modal retrieval. The problem of cross-modal retrieval, for image and text modalities, has been the subject of extensive research in the recent past (Hu et al., 2019; Wang et al., 2019; Wei et al., 2020). Most existing methods are based on one-to-one mapping of instances into a shared embedding space. One popular approach is maximizing correlation between related instances in the embedding space. Rasiwasia et al. (2010) use canonical correlation analysis (CCA) to maximize correlation between images and text. Most recent methods incorporate deep neural networks to learn their embedding models in an end-to-end fashion. Andrew et al. (2013) propose deep CCA (DCCA), and Yan & Mikolajczyk (2015) make it applicable to high dimensional image and text representations cross-modal retrieval. Some attempts have been made to model the bi-directional one-to-many matching. Song & Soleymani (2019) introduced the 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 of correspondences and improve over previous methods based on one-to-one mapping. Li et al. (2019) proposed to use a pretrained object detector (*e.g.* Faster R-CNN (Ren et al., 2015)) to compute region embeddings and build multiple region-word matching. This strategy contributes to significant performance improvements at the expense of a significant increase in computational cost. Instead of deterministic embedding, Chun et al. (2021) proposed to learn stochastic embeddings to address this issue. They embed each instance as a probabilistic distribution rather than a single vector. The probabilistic embeddings can implicitly perform the one-to-many matching between visual and textual concepts. However, none of these methods align the instance-level distributions and the two modalities may still reside in their own spaces, yielding wrong similarity measurement.

Importance Sampling. Importance sampling (Hesterberg, 1988) refers to a collection of Monte Carlo methods where a mathematical expectation with respect to a target distribution is approximated by a weighted average of random draws from another distribution. Importance sampling has been widely applied in domain adaptation (Ben-David et al., 2007; Blitzer et al., 2008; Daume III & Marcu, 2006) to correct the mismatch between the distributions of training and test sets, leading to unbiased estimates of the generalization error (Cortes et al., 2008). To the best of our knowledge, we are the first to apply importance sampling in similarity calibration for probabilistic embeddings.

3 METHOD

We focus on cross-modal retrieval for vision and language data. Let $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$ denote a collection of vision and language data, where $\mathcal{X}$ is a set of images and $\mathcal{Y}$ a set of captions. For a caption $\mathbf{y} \in \mathcal{Y}$ (respectively an image $\mathbf{x} \in \mathcal{X}$), there are multiple matching images $\{\mathbf{x}\}$ (respectively captions $\{\mathbf{y}\}$) due to the multiplicity. The goal is to achieve a shared subspace to measure the similarity between image-caption pairs. Typically, two mapping functions $f_\theta : \mathbf{x} \to \mathbf{u}$ and $g_\phi : \mathbf{y} \to \mathbf{v}$ will be learned where $\mathbf{u}$ and $\mathbf{v}$ are image and caption embeddings, respectively. Rather than point embeddings, $\mathbf{u}$ and $\mathbf{v}$ are probabilistic distributions to facilitate multiplicity. The details of probabilistic embedding will be introduced in Sec. 3.3. For each image-caption pair $(\mathbf{x_i}, \mathbf{y_j})$, we sample their representations as $\mathbf{u_i} \sim p_\theta(\mathbf{u_i} \mid \mathbf{x_i})$ and $\mathbf{v_j} \sim p_\phi(\mathbf{v_j} \mid \mathbf{y_j})$.

In most existing works (Rasiwasia et al., 2010; Hu et al., 2019; Chun et al., 2021), cosine similarity is widely used to measure the similarity:

$$s(I, T) = \mathbf{u_i}^T \mathbf{v_j}. \tag{1}$$

Let $\mathcal{L}_\mathsf{t}$ denotes the image-to-text retrieval loss and $\mathcal{L}_\mathsf{i}$ denotes the text-to-image retrieval loss, hinge-based bidirectional triplet loss (Hermans et al., 2017) can be calculated by:

$$\mathcal{L}_\mathsf{t} = [\alpha - s(I, T) + s(I, \hat{T})]_+, \quad \mathcal{L}_\mathsf{i} = [\alpha - s(I, T) + s(\hat{I}, T)]_+, \tag{2}$$

where $\hat{T} = \text{argmax}_{j \neq T}\, s(I, j)$ and $\hat{I} = \text{argmax}_{i \neq I}\, s(i, T)$ denotes the hardest negatives in a mini-batch. Function $[\cdot]_+ = \max(\cdot, 0)$ and α denotes the margin factor. The overall loss function is:

$$\mathcal{L}_\mathsf{triplet} = \underbrace{\mathbb{E}_{p(\mathbf{u_i})}\mathbb{E}_{p(\mathbf{v_j})} \log \mathcal{L}_\mathsf{t}}_{\textbf{Text retrieval}} + \underbrace{\mathbb{E}_{p(\mathbf{v_j})}\mathbb{E}_{p(\mathbf{u_i})} \log \mathcal{L}_\mathsf{i}}_{\textbf{Image retrieval}}. \tag{3}$$

3.1 SIMILARITY CALIBRATION

We argue that the similarity measurement in Eq. 1 is questionable. The underlining assumption of $\mathbf{u_i}$ and $\mathbf{v_j}$ are well aligned in a shared subspace may not always hold true due to the modality heterogeneity as well as the discrepancy between f_θ and g_ϕ. To address this issue, we propose to calibrate $s(I, T)$ toward more truthful similarity measurement:

$$s(I, T) = (\mathbf{u_i})^T (w \cdot \mathbf{v_j}), \quad where \quad w = \frac{p(\mathbf{u_i})}{p(\mathbf{v_j})}. \tag{4}$$

Let us consider three cases. (1) $w = 1$: The embeddings are well aligned in a share space which degrades to Eq. 1. (2) $w = C$: The misalignment of the embeddings is a constant across all the pairs. (3) $w = \frac{p(\mathbf{u_i})}{p(\mathbf{v_j})}$: The misalignment of the embeddings is a learnable variable where each image-text pair should have a unique w. Obviously, either (1) or (2) is a special (simplified) case of (3).

Now we can calibrate the loss function in Eq. 3 based on the calibrated similarity. More specifically, we maximize the measurement between positive pairs (I, T) by:

$$\begin{aligned} \mathbb{E}_{p(\mathbf{u_i})}\mathbb{E}_{p(\mathbf{v_j})} \log s(I, T) &= \mathbb{E}_{p(\mathbf{u_i})}[\mathbb{E}_{p(\mathbf{v_j})} \log(\mathbf{u_i}^T \mathbf{v_j}) + \mathbb{E}_{p(\mathbf{v_j})} \log \frac{p(\mathbf{u_i})}{p(\mathbf{v_j})}] \\ &= \mathbb{E}_{p(\mathbf{u_i})}[\underbrace{\mathbb{E}_{p(\mathbf{v_j})} \log(\mathbf{u_i}^T \mathbf{v_j})}_{\textbf{Point matching}} - \underbrace{\text{KL}(p(\mathbf{v_j}) || p(\mathbf{u_i}))}_{\textbf{Distribution matching}}]. \end{aligned} \tag{5}$$

At the same time, we minimize the measurement between negative pairs $(I, \hat{T})$ or $(\hat{I}, T)$ by: $\mathbb{E}_{p(\mathbf{u_i})}[\mathbb{E}_{p(\mathbf{v_j})} \log(\mathbf{u_i}^T \mathbf{v_j}) + \text{KL}(p(\mathbf{v_j}) || p(\mathbf{u_i}))]$. We highlight that Eq. 5 is the combination of *point matching* and *distribution matching*. The proposed calibration can effectively align two embeddings and enlarge the aligned subspace (Fig. 1 (c)), facilitating accurate one-to-many matching.

3.2 PRE-TRAINING TASKS FOR LANGUAGE MODELS

Pre-training tasks play the key role for downstream tasks. Although pre-training tasks have been intensively studied in image (Zbontar et al., 2021; He et al., 2020; Chen et al., 2020b) and text representations (Devlin et al., 2019; Karpukhin et al., 2020; Gao et al., 2021; Raffel et al., 2020) individually,

they are seldom investigated in cross-modal retrieval. To investigate which pre-training task can facilitate better caption embeddings, we evaluate three pre-training tasks for language models: Masked Language Modelling (MLM) (Devlin et al., 2019), Dense Passage Retrieval (DPR) (Karpukhin et al., 2020), and Simple Contrastive Sentence Embedding (SimCSE) (Gao et al., 2021).

MLM refers to randomly mask some input tokens and predict these masked tokens. In Devlin et al. (2019), 15% of all the tokens in a sequence are randomly masked. To mitigate the mismatch between pre-training and fine-tuning, "masked" words are not always replaced by the actual [MASK] token. In detail, if the token is chosen to be masked, it will be replaced by the [MASK] token with an 80% probability and a random token with a 10% probability, and keep unchanged with a 10% probability.

DPR is designed for open-domain question answering. The goal is to retrieve the relevant passages including the answers for the question. To achieve it, two encoders are employed to get the question and passage embeddings. A contrastive learning framework is utilized to maximize the inner product between positive pairs (the question and relevant passage vectors) and minimize it between negative pairs (the question and irrelevant passage vectors).

SimCSE leverages a simple contrastive learning framework to learn sentence embedding. They propose an unsupervised and a supervised approach. In unsupervised approach, the model predicts the input sentence itself in a contrastive framework where standard dropout is used to generate positive pairs. In supervised approach, they leverage annotated pairs into a contrastive learning framework where"entailment" (manually created sentences with the same semantics) are positives and "contradiction" (manually created sentences with the opposite semantics) are hard negatives. SimCSE greatly advances the state-of-the-art on standard on semantic textual similarity tasks.

In Sec. 4.4, we empirically demonstrate that SimCSE outperforms the other two pre-training tasks, possibly due to that the contrastive learning with *entailment* and *contradiction* contribute to better caption embeddings for cross-modal retrieval.

3.3 Implementation

We apply image and text augmentation to approximate $p(\mathbf{u_i})$ and $p(\mathbf{v_j})$, respectively. KL divergence in Eq. 5 is asymmetric and well-known for the problem of vanishing gradient. To address it, we use Wasserstein distance (Villani, 2003) to minimize the discrepancy between $p(\mathbf{u_i})$ and $p(\mathbf{v_j})$, yielding more stable *distribution matching*.

Data augmentation. For image augmentation, we assume access to a set Ψ, where each element is a specific image transformation. We can generate multiple augmentations by sampling from Ψ. For text augmentation, we use a fine-tuned T5 model (Raffel et al., 2020) to get multiple paraphrased captions from the original caption.

Wasserstein distance. We use 2-Wasserstein ($\mathcal{W}_2$) distance to align the distribution between $p(\mathbf{u_i})$ and $p(\mathbf{v_j})$. The $\mathcal{W}_2$ coupling distance between $p(\mathbf{u_i})$ and $p(\mathbf{v_j})$ on $\mathbb{R}^n$ is:

$$\mathcal{W}_2(p(\mathbf{u_i}); p(\mathbf{v_j})) := \inf \mathbb{E}\left(\|\mathbf{u_i} - \mathbf{v_j}\|_2^2\right)^{1/2}, \tag{6}$$

where the infimum runs over all random vectors $(\mathbf{u_i}, \mathbf{v_j})$ of $\mathbb{R}^n \times \mathbb{R}^n$ with $\mathbf{u_i} \sim p(\mathbf{u_i})$ and $\mathbf{v_j} \sim p(\mathbf{v_j})$. We assume $\mathbf{u_i}$ and $\mathbf{v_j}$ follow Gaussian distribution in the embedding space: $p(\mathbf{u_i}) = \mathcal{N}(\mu_i, \Sigma_i)$, $p(\mathbf{v_j}) = \mathcal{N}(\mu_j, \Sigma_j)$. Eq. 6 is reduced to:

$$\mathcal{W}_2(p(\mathbf{u_i}); p(\mathbf{v_j}))^2 = \|\mu_i - \mu_j\|_2^2 + \left\|\Sigma_i^{1/2} - \Sigma_j^{1/2}\right\|_2^2. \tag{7}$$

We empirically demonstrate that 2-Wasserstein distance outperforms KL divergence by a large margin in Sec. 4.4.

4 Experiments

To evaluate the effectiveness of the proposed similarity calibration, we perform experiments in both image-to-text and text-to-image retrieval on two widely used datasets. We perform ablation studies to investigate the key components of our approach. We also compare with previous state-of-the-art methods for cross-modal retrieval.

4.1 DATASETS AND METRICS

Datasets. We conducted experiments on the two widely-used benchmark datasets: Flickr30K (Young et al., 2014) and MS-COCO (Lin et al., 2014), to evaluate our proposed model and compare with several state-of-the-art baselines. **Flickr30K**. This dataset consists of 31,783 images. Each image is described by 5 different sentences. Following the settings in previous work (Chen & Luo, 2020; Lee et al., 2018; Qu et al., 2020), this dataset is split into 29,783 training images, 1,000 validation images, and 1,000 testing images. **MS-COCO**. It is a large-scale dataset including 123,287 images, where each image is associated with 5 annotated sentences. Similarly, we followed the split of (Chen & Luo, 2020; Lee et al., 2018; Qu et al., 2020), *i.e.*, 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. We report the results on the two evaluation settings: 1) MS-COCO 1K, the final result is calculated by averaging the results over 5-folds of 1K testing images; and 2) MS-COCO 5K, the evaluation result is directly calculated on the full 5K testing images.

Metrics. **Recall@k** (R@k). We report rank-1 (R@1), rank-5 (R@5), and rank-10 (R@10) results on both datasets. **Recall-Precision** (R-P). Musgrave et al. (2020) proposed the Recall-Precision (R-P) metric as an alternative. For each query, we compute the ratio between matched items and top-r retrieved items, where r is the number of ground-truth matches. This precision metric has a promising property that a retrieval model achieves the highest R-P score if and only if it retrieves all the matched items before the negatives. Compared to Recall@k, R-P score can better evaluate the one-to-many matching for image-to-text retrieval.

Method	Backbone	Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10
DAN (Nam et al., 2017)	VGG-19	41.4	73.5	82.5	31.8	61.7	72.5
RRF-Net (Liu et al., 2017)	ResNet-152	47.6	77.4	87.1	35.4	68.3	79.9
CMPM +CMPC (Zhang & Lu, 2018)	ResNet-152	49.6	76.8	86.1	37.3	65.7	75.5
DAN (Nam et al., 2017)	ResNet-152	55.0	81.8	89.0	39.4	69.2	79.1
NAR (Liu et al., 2019)	ResNet-152	55.1	80.3	89.6	39.4	68.8	79.9
VSE++ (Faghri et al., 2016)	ResNet-152	52.9	80.5	87.2	39.6	70.1	79.5
SCO (Huang et al., 2018)	ResNet-152	55.5	82.0	89.3	41.1	70.5	80.1
GXN (Gu et al., 2018)	ResNet-152	56.8	–	89.6	41.5	–	80.1
TIMAM (Sarafianos et al., 2019)	ResNet-152	53.1	78.8	87.6	42.6	71.6	81.9
Ours	ResNet-152	**62.7**	**86.2**	**91.5**	**49.6**	**78.8**	**85.2**

Table 1: Comparison (%) on Flickr30K (Young et al., 2014). Our method outperforms all baselines by a large margin in both *Image-to-Text* and *Text-to-Image* retrieval.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA (Karpathy & Fei-Fei, 2015)	16.5	39.2	52.0	10.7	29.6	42.2
GMM-FV (Klein et al., 2015)	17.3	39.0	50.2	10.8	28.3	40.1
Order (Vendrov et al., 2016)	23.3	–	65.0	18.0	–	57.6
VQA-A (Lin & Parikh, 2016)	23.5	50.7	63.6	16.7	40.5	53.8
CMPM (Zhang & Lu, 2018)	31.1	60.7	73.9	22.9	50.2	63.8
VSE++ (Faghri et al., 2016)	41.3	71.1	81.2	30.3	59.4	72.4
SCO (Huang et al., 2018)	42.8	72.3	83.0	<u>33.1</u>	62.9	<u>75.5</u>
PVSE (K=1) (Song & Soleymani, 2019)	41.7	73.0	83.0	30.6	61.4	73.6
PVSE (Song & Soleymani, 2019)	**45.2**	<u>74.3</u>	<u>84.5</u>	32.4	<u>63.0</u>	75.0
PCME (Chun et al., 2021)	44.2	73.8	83.6	31.9	62.1	74.5
Ours	<u>44.8</u>	**75.1**	**84.8**	**33.8**	**64.8**	**76.4**

Table 2: Comparison (%) on MS-COCO (Lin et al., 2014) 5K test set . Our approach achieves the best results in *Text-to-Image* retrieval on the 5K test set.

4.2 IMPLEMENTATION DETAILS

We use ResNet152 (He et al., 2016) and BERT (Devlin et al., 2019) as image and text encoder, respectively. We use Adam optimizer (Kingma & Ba, 2014) for model training with the mini-batch

size of 64 and the epoches of 30. We set the initial learning rate as 0.0002 with decaying 10% of every 15 epochs. The dimension of visual features is 2,048. The basic version of the pre-trained BERT (Devlin et al., 2019) is utilized, equipped with 12 layers, 12 heads, 768 hidden units, and 110M parameters in total, to get the word embeddings with dimension of 768. We set the dimension of joint embedding space as 512.

4.3 Retrieval Comparison

Results on Flickr30K (Young et al., 2014): We compare our approach to previous state-of-the art methods. We report the results on the Flickr30K (Young et al., 2014) in Tab. 1. Similar to the most methods, we use ResNet-152 He et al. (2016) as the image encoder for fair comparison. Our method outperforms all methods by a large margin in both *Image-to-Text* and *Text-to-Image* retrieval. In detail, our method outperforms the second best by 5.9% and 7.0% in *Image-to-Text* and *Text-to-Image* retrieval, respectively. Specially, TIMAM (Sarafianos et al., 2019) also used Bert (Kenton & Toutanova, 2019) as the backbone for text encoder. TIMAM utilizes adversarial training to align the distributions between the two modalities through adversarial training (Goodfellow et al., 2015) and results in modality-invariant embeddings. The superior performance demonstrates that the proposed similarity calibration can facilitate more accurate similarity measurement and correctly capture the multiplicity of correspondences.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA (Karpathy & Fei-Fei, 2015)	38.4	69.9	80.5	27.4	60.2	74.8
GMM-FV (Klein et al., 2015)	39.4	67.9	80.9	25.1	59.8	76.6
m-CNN (Ma et al., 2015)	42.8	73.1	84.1	32.6	68.6	82.8
Order (Vendrov et al., 2016)	46.7	–	88.9	37.9	–	85.9
DSPE (Wang et al., 2016)	50.1	79.7	89.2	39.6	75.2	86.9
VQA-A (Lin & Parikh, 2016)	50.5	80.1	89.7	37.0	70.9	82.9
2WayNet (Eisenschtat & Wolf, 2017)	55.8	75.2	–	39.7	63.3	–
RRF-Net (Liu et al., 2017)	56.4	85.3	91.5	43.9	78.1	88.6
CMPM (Zhang & Lu, 2018)	56.1	86.3	92.9	44.6	78.8	89.0
VSE++ (Faghri et al., 2016)	64.6	90.0	95.7	52.0	84.3	92.0
GXN (Gu et al., 2018)	68.5	–	97.9	56.6	–	94.5
SCO (Huang et al., 2018)	**69.9**	92.9	97.5	56.7	87.5	94.8
PVSE (K=1) (Song & Soleymani, 2019)	66.7	91.0	96.2	53.5	85.1	92.7
PVSE (Song & Soleymani, 2019)	69.2	91.6	96.6	55.2	86.5	93.7
PCME (Chun et al., 2021)	68.8	91.6	96.7	54.6	86.3	93.8
Ours	69.7	**93.5**	97.8	**57.0**	**88.2**	**95.6**

Table 3: Comparison (%) on MS-COCO (Lin et al., 2014) 1K test set.

Results on MS-COCO (Lin et al., 2014): We compare our approach to previous state-of-the art methods. We report the results on 5k and 1k test sets in Tabs 2 and 3. Our approach outperforms most of the baselines, and achieves the new state-of-the-art on the *Text-to-Image* on the 5K test set. We note that both GXN (Gu et al., 2018) and SCO (Huang et al., 2018) are trained with multiple tasks apart from the image-text matching: GXN (Gu et al., 2018) conducts cross-modal synthesis during model training, while SCO (Huang et al., 2018) jointly classifies the objects and their orders during model training. Our model is only trained with the image-text matching and yield competitive results. We also compare our approach to other methods tailored for one-to-many matching: VSE++ (Faghri et al., 2016), PVSE (Song & Soleymani, 2019), VSRN (Li et al., 2019), AOQ (Chen et al., 2020a), and PCME (Chun et al., 2021). Specially, PCME (Chun et al., 2021) is the the first work that uses probabilistic embeddings for cross-modal retrieval. In the 1K test set, our method outperforms all of these methods. In detail, our method outperforms PCME (Chun et al., 2021) by 0.95% and 2.4% in *Image-to-Text* and *Text-to-Image* retrieval, respectively. The results demonstrate that our method is capable of better capturing the one-to-many matching in both directions.

4.4 Ablation Studies

Method	Flickr30K					MS-COCO				
	Image-to-Text			Text-to-Image		Image-to-Text			Text-to-Image	
	R@1	R@5	R-P	R@1	R@5	R@1	R@5	R-P	R@1	R@5
Ours	**62.7**	**86.2**	**47.2**	**49.6**	**78.8**	**44.8**	**75.1**	**32.5**	33.8	**64.8**
w/o Calibration	60.7	85.2	45.6	48.0	77.4	42.4	72.8	31.3	32.6	63.5
KL Divergence	61.0	85.4	46.5	47.8	77.0	43.5	73.7	31.9	33.2	64.2
JS Divergence	61.6	86.0	46.4	48.6	78.1	43.9	74.5	32.1	**34.0**	64.5

Table 4: Ablation study of similarity calibration on Flickr30K (Young et al., 2014) and MS-COCO (Lin et al., 2014). The proposed similarity calibration significantly improves both *Image-to-Text* and *Text-to-Image* retrieval.

Validation of similarity calibration: The *similarity calibration* in Eq. 5 plays the key role in calibrating the similarity measurement and enlarging the shared representation space. We implement three variants for comparison: 1) Without Calibration: the model is only trained with *Point Matching* as shown in Eq. 5. 2) Kullback–Leibler (KL) divergence. 3) Jensen-Shannon (JS) divergence : $\text{JSD}(p(\mathbf{u_i}), p(\mathbf{v_j})) = \frac{1}{2}[\text{KL}(p(\mathbf{u_i}), p(\mathbf{v_j})) + \text{KL}(p(\mathbf{v_j}), p(\mathbf{u_i}))]$. Results on Flickr30K (Young et al., 2014) and MS-COCO (Lin et al., 2014) are reported in Tab. 4. Our approach outperforms the other three variants in Flickr30K (Young et al., 2014) across all evaluation metrics. In *Image-to-Text* retrieval on Flickr30K (Young et al., 2014), our method outperforms *w/o Calibration* and *JSD* by 2% and 1.1% in R@1. In R-P, our method outperforms *w/o Calibration* and *JSD* by 1.6% and 0.8%. In *Text-to-Image* retrieval on Flickr30K (Young et al., 2014), our method outperforms *w/o Calibration* and *JSD* by 1.6% and 1.0% in R@1. In *Image-to-Text* retrieval on MS-COCO (Lin et al., 2014), our method outperforms *w/o Calibration* and *JSD* by 2.4% and 0.9% in R@1. In R-P, our method outperforms *w/o Calibration* and *JSD* by 1.2% and 0.4%. In *Text-to-Image* retrieval on MS-COCO (Lin et al., 2014), although the result of R@1 is slightly lower than that of *JSD*, our method outperforms *w/o Calibration* and *JSD* by 1.3% and 0.3% in R@5. To further evaluate the effectiveness of the proposed similarity calibration, we conduct experiments with limited training data. Results of R-P on Flickr30K (Young et al., 2014) are shown in Fig. 2. The improvements are more significant especially when the training data are extremely limited (38.3% v.s. 34.4% with 20% of training pairs). Results demonstrate the effectiveness of the proposed similarity calibration.

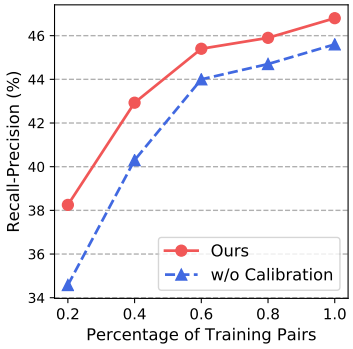


Figure 2: Evaluation of similarity calibration on limited training data.

Validation of data augmentation: We apply image and text augmentation to approximate the distributions of the two modalities. To better understand the effect of augmentation, we conduct experiments by varying the number of augmentations. In detail, we vary the number of image augmentations from 1 to 10 for each image, and vary the number of text augmentations from 1 to 5 for each caption. Note that, in both Flickr30K (Young et al., 2014) and MS-COCO (Lin et al., 2014), we have 5 captions for each image. As a result, there will be 30 captions for each image if we create 5 augmentations for each caption. We report the results of Recall-Precision on different number of augmentations in Fig. 3. In both image and caption augmentations, more augmentations can consistently improve R-P due to more accurate distribution distribution estimation.

Evaluation of pre-training tasks: To investigate which pre-training task of language models can boost the performance for cross-modal retrieval, we evaluate three pre-training tasks for language models: Masked Language Modelling (MLM) (Devlin et al., 2019), Dense Passage Retrieval (DPR) (Karpukhin et al., 2020), and Simple Contrastive Sentence Embedding (SimCSE) (Gao et al., 2021). In SimCSE, we evaluate the two variants: unsupervised (un) and supervised (su) approaches. In unsupervised approach, the model predicts the input sentence itself in a contrastive objective, with only standard dropout as noise. In supervised approach, they leverage annotated pairs in a contrastive learning framework by using "entailment" (manually created sentences with the same semantics) pairs as positives and "contradiction" (manually created sentences with the opposite

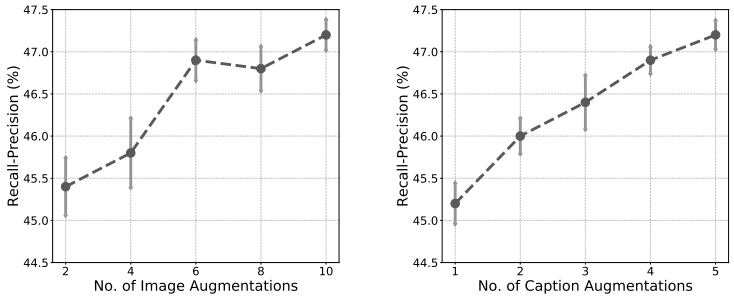


Figure 3: Ablation study on the number of data augmentations. More augmentations contribute to higher Recall-Precision due to more accurate distribution estimation.

Task	Image-to-Text			Text-to-Image	
	R@1	R@5	R-P	R@1	R@5
MLM (Devlin et al., 2019)	60.3	85.5	45.4	48.0	76.9
DPR (Karpukhin et al., 2020)	59.7	84.9	45.6	47.2	77.6
SimCSE (un) (Gao et al., 2021)	60.1	85.3	45.6	48.3	77.1
SimCSE (su) (Gao et al., 2021)	**62.7**	**86.2**	**47.2**	**49.6**	**78.8**

Table 5: Comparison (%) of pre-training tasks on Flickr30K (Young et al., 2014). SimCSE (su) outperforms other tasks in both *Image-to-Text* and *Text-to-Image* retrieval.

semantics) pairs as hard negatives. Results on Flickr30K (Young et al., 2014) are reported in Tab. 5. As shown in Tab. 5, SimCSE (su) outperforms others in both *Image-to-Text* and *Text-to-Image*. In comparison to MLM, SimCSE (su) can not only boost the performance (2.4% in R@1 and 1.8% in R-P) in *Text-to-Image*, but also boost the performance (1.6% in R@1) in *Image-to-Text*. Results demonstrate that contrastive learning with *entailment* and *contradiction* contribute to better sentence embedding.

5 Conclusion

In this paper, we introduced a novel calibration method for probabilistic embeddings in cross-modal retrieval. We estimate the density ratio between the distributions of the two modalities, and use it to calibrate the similarity measurement in the embedding space. The similarity calibration can effectively align the distributions between the two modalities and enlarge the shared representation space, facilitating more accurate one-to-many matching. In addition, we further evaluate three pre-training tasks of language models for cross-modal retrieval. We empirically found that SimCSE outperforms the other two pre-training tasks due to better sentence embeddings. Extensive experiments as well as ablation studies on two benchmarks demonstrate its superior performance in tackling the multiplicity of cross-modal retrieval.

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pp. 1247–1255. PMLR, 2013.

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. 2008.

Tianlang Chen and Jiebo Luo. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10583–10590, 2020.

Tianlang Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. In *European Conference on Computer Vision*, pp. 549–565. Springer, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8415–8424, June 2021.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pp. 38–53. Springer, 2008.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4601–4611, 2017.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, 2016.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7181–7189, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

Timothy Classen Hesterberg. *Advances in importance sampling*. PhD thesis, Stanford University, 1988.

Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pp. 635–644, 2019.

Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6163–6171, 2018.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *arXiv:1412.6980 [cs.LG]*, 2014.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4437–4446, 2015.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 201–216, 2018.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4654–4662, 2019.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, pp. 261–277. Springer, 2016.

Chunxiao Liu, Zhendong Mao, Wenyu Zang, and Bin Wang. A neighbor-aware approach for image-text matching. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3970–3974. IEEE, 2019.

Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4107–4116, 2017.

Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pp. 2623–2631, 2015.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision*, pp. 681–699. Springer, 2020.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 299–307, 2017.

Seong Joon Oh, Kevin P Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C Gallagher. Modeling uncertainty with hedged instance embeddings. In *International Conference on Learning Representations*, 2018.

Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. Context-aware multi-view summarization network for image-text matching. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1047–1055, 2020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 251–260, 2010.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99, 2015.

Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5814–5824, 2019.

Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1979–1988, 2019.

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

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5005–5013, 2016.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5764–5773, 2019.

Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10941–10950, 2020.

Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3441–3450, 2015.

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

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 686–701, 2018.