TRUST-CONSISTENT VISUAL SEMANTIC EMBEDDING FOR IMAGE-TEXT MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual Semantic Embedding (VSE), as a link between Computer Vision and Natural Language Processing, aims at jointly learning cross-modal embeddings to bridge the discrepancy across visual and textual spaces. In recent years, VSE has achieved great success in image-text matching benefiting from the outstanding representation power of deep learning. However, existing methods produce retrieved results only relying on the ranking of cross-modal similarities, even if the retrieved results are unreliable and uncertain. That is to say, they cannot self-evaluate the quality of retrieved results for trustworthy retrieval, resulting in ignoring the ubiquitous uncertainty in data and models. To address this problem, we propose a novel VSE-based method for image-text matching, namely Trust-consistent Visual Semantic Embedding (TcVSE), to embrace trustworthy retrieval and self-evaluation for image-text matching. To be specific, first, TcVSE models the evidence based on cross-modal similarities to capture accurate uncertainty. Second, a simple yet effective consistency module is presented to enforce subjective opinions of bidirectional VSE models (*i2t*+*t2i*) to be consistent for high reliability and accuracy. Finally, extensive comparison experiments are conducted to demonstrate the superiority of TcVSE on two widely-used benchmark datasets, i.e., Flickr30K and MS-COCO. Furthermore, some qualitative experiments are carried out to provide comprehensive and insightful analyses for the reliability and rationality of our method.

1 INTRODUCTION

Visual Semantic Embedding aims to learn a shared embedding space to enforce visual data coincide with their corresponding semantic textual descriptions, which is an important approach to understanding the cross-modal semantic association for downstream applications, such as image-text matching Faghri et al. (2017) and visual question-answering Malinowski et al. (2015), etc. Thus, the key issue of VSE is how to eliminate the discrepancy across images and texts to learn a reliable common embedding space. To address this issue, numerous methods attempt to project visual and textual data into a latent common space. However, it is still unknown to self-evaluate the retrieval performance to achieve interpretable and reliable inference.

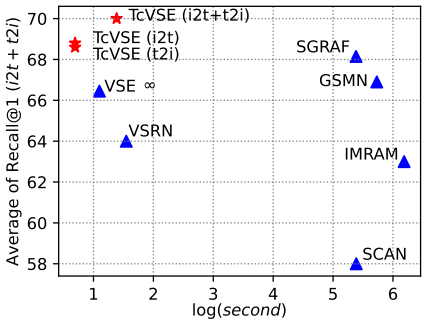


Figure 1: Performance comparison on image-text matching. We show the average Recall@1 vs. the time cost of cross-modal retrieval on Flickr30K. Our method, TcVSE, is shown in red.

In this paper, we focus on image-text matching (ITM), one of the fundamental tasks of cross-modal learning, i.e., cross-modal retrieval, which expects to search the most relevant sentences for a given image query (*i2t*) or retrieve the related images from a given sentence query (*t2i*) according to the pairwise visual-semantic similarities. Some early works based on VSE Kiros et al. (2014); Wang et al. (2016); Faghri et al. (2017) leverage the powerful feature extraction capability of deep neural networks (DNNs) to obtain the global representation of images and texts, such as VGG Simonyan & Zisserman (2014), ResNet He et al. (2016), and GRU Chung et al. (2014), etc., by maximizing the correlated cross-modal similarities. More granularly, recent

VSRN Li et al. (2019) performs reasoning with Graph Convolutional Neural networks (GCNs) Kipf & Welling (2016) to generate enhanced visual representations, which captures both objects and corresponding semantic relationships for better visual semantic embedding. VSE∞ Chen et al. (2021) presents an adaptive pooling strategy (GPO) that aggregates (region-based or grid-based) local features to lean a better common representation. Unlike the aforementioned VSE-based methods, some works Lee et al. (2018); Chen et al. (2020); Wu et al. (2019); Liu et al. (2020); Diao et al. (2021); Cheng et al. (2022); Li et al. (2022a) present a specific mechanism or model to explicitly learn and integrate the fine-grained relationships between image regions and word tokens for cross-modal similarity inference.

Although prior approaches could achieve promising performance, they are only able to estimate image-text similarities for cross-modal retrieval, wherein image-text pairs with high similarity are taken for granted as matched. Since the ubiquitous uncertainty in data and models, it is inevitable to produce unreliable retrieval results. Therefore, it requires revisiting the questions such as "*Is this retrieval trustworthy?*" to evaluate the uncertainty or unreliability of predictions. To this end, it is valuable and necessary to measure such uncertainty for self-evaluation, but less touched in existing image-text matching methods.

To address this problem, we propose a novel VSE framework, termed Trust-consistent Visual Semantic Embedding (TcVSE). Not only does TcVSE outperform prior works (Figure 1), but it is also more efficient, achieving trustworthy image-text matching. More specifically, **(1)** we employ Evidential Deep Learning (EDL) built on the Dempster-Shafer Theory of Evidence (DST) Yager & Liu (2008) and the Subjective Logical Theory Sensoy et al. (2018) (SL) into VSE models to capture the uncertainty, thus endowing the model with the ability to self-evaluate retrieval quality. Following the principles of DST and SL, we consider the pairwise similarity measured by VSE as a source of evidence and parameterize the evidence as a Dirichlet distribution, which not only models the density of query probabilities but also the uncertainty. **(2)** Unlike prior EDL methods, our TcVSE focuses on ITM instead of classification. Thus, our TcVSE should overcome two challenges to apply EDL on ITM, namely instance retrieval and bidirectional inference. To tackle the first challenge, we relax the instance-level retrieval to a K-way querying for training, thus enabling uncertainty estimation via cross-modal similarity. To counter the second challenge, two VSE branches (*i2t* and *t2i*) with EDL are proposed to learn bidirectional retrieval, however, the difference between the two tasks unavoidably leads to the gap between their uncertainty. To address the problem, we present a simple yet effective consistency module to enforce subjective opinions of different branches to be consistent for more reliable uncertainty estimation, thus embracing performance improvement. **(3)** Finally, we demonstrate the effectiveness and superiority of our method with extensive experiments on two widely used benchmark datasets, i.e., Flickr30K and MS-COCO. The comprehensive ablation studies and insightful analyses verify the reliability and practicability of our method.

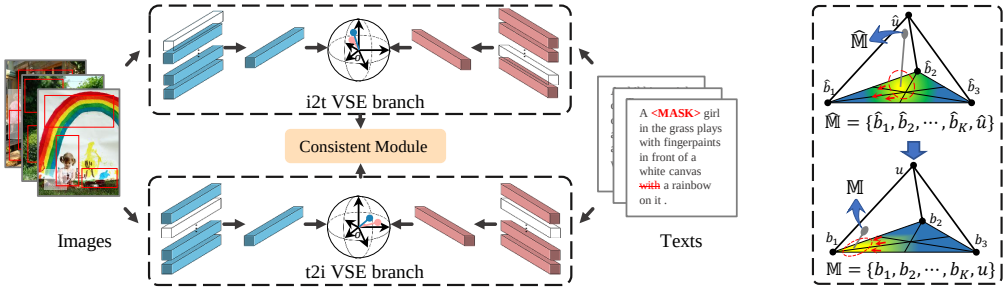
2 Trust-consistent Visual Semantic Embedding

In this section, we summarize our method in Section 2.1 and elaborate on how to estimate the evidence-based uncertainty for trustworthy image-text matching in Section 2.2. Moreover, we present a Consistent Module to make two VSE branches obtain consistent predictions on subjective opinions during evidential deep learning in Section 2.3.

2.1 Overview

To achieve trustworthy image-text matching, unlike most standard methods, TcVSE utilizes EDL and a consistent module to accurately measure the visual-textual similarity and additionally quantify the uncertainty of the VSE model for self-evaluation. Figure 2 shows the framework of our proposed method. We first define our Visual Semantic Embedding model for image-text matching as illustrated in Figure 2(a). Let $(\mathcal{V}, \mathcal{C})$ denote a visual and textual dataset, which contains a set of images $\mathcal{V}$ and a set of texts $\mathcal{C}$.

Feature Encoding: For any sample pair $(\boldsymbol{u}, \boldsymbol{c})$ in $(\mathcal{V}, \mathcal{C})$, their feature representations could be encoded by some deep backbone networks, e.g., Faster-RCNN for visual features and Bi-GRU for



(a) Trust-consistent Visual Semantic Embedding (TcVSE) (b) Consistent Module

Figure 2: Overview of the proposed approach. (a) shows the pipeline of our TcVSE, which consists of two independent bidirectional VSE models, i.e., $i2t$ and $t2i$ VSE branches. Notably, each VSE branch could be used to measure similarities for bidirectional retrieval (sentence retrieval given image query and image retrieval given sentence query). Specifically, each VSE branch contains a visual backbone (e.g., Faster-RCNN) and a textual backbone (e.g., Bi-GRU or Bert-base). First, in each VSE branch, an image or text will be fed into the corresponding backbone to extract fine-grained features. Second, *Max pooling* is conducted to aggregate the fine-grained feature vectors for similarity calculations. Third, a novel cross-modal evidential learning is applied in TcVSE to optimize the branches to capture the uncertainty lurking in the obtained similarities for trustworthy retrieval. Finally, our Consistent Module (shown in (b)) enforces the predicted subjective opinions of the two branches to be more consistent for a reliable uncertainty estimation.

textual features, respectively:

$$\mathrm{V}(\boldsymbol{u}, \Theta_\phi) : \boldsymbol{u} \rightarrow \{\boldsymbol{x}_i\}_{i=1}^{V}, \boldsymbol{x}_i \in \mathbb{R}^d,$$
$$\mathrm{T}(\boldsymbol{c}, \Theta_\psi) : \boldsymbol{c} \rightarrow \{\boldsymbol{r}_j\}_{j=1}^{M}, \boldsymbol{r}_j \in \mathbb{R}^d$$

where d is the dimensionality of the joint embedding space, $\mathrm{V}(*, \Theta_\phi)$ and $\mathrm{T}(*, \Theta_\psi)$ are respectively visual and textual backbones, Θ_ϕ and Θ_ψ are respectively the corresponding model parameters, $\{\boldsymbol{x}_i\}_{i=1}^{V}$ is a set of V encoded local region features, $\{\boldsymbol{r}_j\}_{j=1}^{M}$ is a set of M word token features, $\boldsymbol{r}_j \in \mathbb{R}^d$ is a word token feature and M is the number of words for $\boldsymbol{c}$. Following Hüllermeier & Waegeman (2021), we randomly discard the region features extracted by the backbone network (Faster-RCNN) to achieve data augmentation, which is different from the common augmentation of the raw image, e.g., cropping, rotation, etc. Meanwhile, "Mask", "Discard", or "Swap" operations are performed on the word tokens for text augmentation.

Similarity Representation: To obtain the global similarity, the encoded visual features $\{\boldsymbol{x}_i\}_{i=1}^{N}$ and textual features $\{\boldsymbol{r}_j\}_{j=1}^{M}$ would be aggregated by *Max pooling* into a common embedding space.

$$\boldsymbol{v} = \mathrm{MaxPooling}\left(\{\boldsymbol{x}_i\}_{i=1}^{N}\right), \quad \boldsymbol{t} = \mathrm{MaxPooling}\left(\{\boldsymbol{r}_j\}_{j=1}^{M}\right).$$

Then, the similarity score of $(\boldsymbol{v}, \boldsymbol{t})$ is measured by the cosine similarity as follows:

$$\mathrm{S}(\boldsymbol{v}, \boldsymbol{t}) = \frac{\boldsymbol{v}^\top \boldsymbol{t}}{\|\boldsymbol{v}\| \cdot \|\boldsymbol{t}\|}. \tag{1}$$

Learning with TcVSE: A VSE model aims at minimizing the visual-semantic distance in a common space, i.e., maximizing the similarity of matched visual and textual samples. Our TcVSE aims to achieve that goal while also endowing the VSE models with the reliable capability of uncertainty estimation. More specifically, TcVSE conducts a two-step learning process to optimize models. The first step is to optimize the uncertainty-aware loss $\mathcal{L}_u$ based on the cross-modal evidential deep learning. The second step is multiple optimizations for opinion-based consistency loss $\mathcal{L}_c$. See Algorithm 1 for more details on the optimization process.

2.2 Uncertainty Estimation

In this section, we follow the notions and principles of evidential deep learning (EDL) Sensoy et al. (2018) to model the uncertainty of VSE models. To estimate uncertainty, the Dempster-Shafer

Theory of Evidence (DST) Yager & Liu (2008) and the theory of Subjective Local (SL) Jsang (2016) are employed to build the learning paradigm of EDL. The existing EDL learns a deterministic model from the observable evidence supporting subjective opinions (i.e., model predictions). However, these methods almost focus on unimodal classification and less touching image-text matching.

For image-text matching, VSE projects the visual and textual feature representations into a common space, thus making it possible to measure the similarity across different modalities. Different from existing EDL methods Sensoy et al. (2018), VSE does not have a nonlinear classifier to predict the evidence, thus making it difficult to quantify the uncertainty. To address the issue, our TcVSE relaxes the instance-level retrieval to a K-way querying, thus the evidences could be estimated by the cross-modal similarities, i.e., $\boldsymbol{e}_k = [g(s_{k1}), g(s_{k2}), \cdots, g(s_{kK})]$ for the k-th query, where K is the number of matching events and $g(\cdot)$ is a function to transform similarity into a non-negative evidence (i.e., $e \in [0, +\infty)$) as bellow:

$$e = g(s) = \text{ReLU}(s/\tau) \text{ or } \exp(s/\tau), \tag{2}$$

where s is the visual-semantic similarity computed by Equation (1) and $0 < \tau < 1$ is a temperature parameter Wu et al. (2018). To model the uncertainty, the similarity-based evidence vector $\boldsymbol{e}$ could be associated with the parameters of a Dirichlet distribution $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_K]$ ($\alpha_k = e_k + 1$) built on SL theory, which provides an overall uncertainty mass u and a belief mass b_i for each singleton that is one of K retrieval events of a *Qurey* in image-text matching. These $K+1$ masses are defined as

$$b_k = \frac{e_k}{S} = \frac{\alpha_k - 1}{S} \text{ and } u = \frac{K}{S}, \tag{3}$$

where $S = \sum_{k=1}^{K} (e_k + 1) = \sum_{k=1}^{K} \alpha_k$ and $\sum_{k=1}^{K} b_k + u = 1$. The belief masses $\boldsymbol{b} = [b_1, b2, \cdots, b_K]$ could be treated as subjective opinions corresponding to the parameters of Dirichlet distribution $\boldsymbol{\alpha}$ and the S could be considered as the distribution strength.

Intuitively, ITM could be viewed as a process of retrieving counterparts with the highest matching probability from different modalities. Hence, the matching probability assignment over the retrieved samples of each "*Query*" could be denoted as $\mathbf{p} = [p_1, p_2, \cdots, p_K]$, where $\sum_{i=1}^{K} p_i = 1$. By using the Dirichlet distribution to model such probability assignment, given an opinion, the expected probability of the k-th matched event can written as $\mathbb{E}_{D(\mathbf{p}|\boldsymbol{\alpha})}[p_k] = \int p_k D(\mathbf{p} \mid \boldsymbol{\alpha}) d\mathbf{p} = \frac{\alpha_k}{S}$, where the Dirichlet distribution with parameters $\langle \alpha_1, \alpha_2, \cdots, \alpha_K \rangle$ parameterized over the evidence $\langle e_1, e_2, \cdots, e_K \rangle$ expresses the density of such probability assignment and simultaneously models the overall uncertainty Jsang (2016). The density function is given by

$$D(\mathbf{p} \mid \boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{K} p_j^{\alpha_j - 1} & \text{for } \mathbf{p} \in \mathcal{S}_K \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $B(\boldsymbol{\alpha})$ is the K-dimensional multinomial *beta* function and S_K is the K-dimensional unit simplex. For a deep classifier, the widely used loss function is cross-entropy, formally as

$$\mathcal{L}_{ce}(\boldsymbol{y}, \mathbf{p}) = -\sum_{j=1}^{K} y_j \log(p_j).$$

Considering the density function $D(\mathbf{p} \mid \boldsymbol{\alpha})$ molded by the Dirichlet distribution $\boldsymbol{\alpha}$, the *Bayes risk* of $\mathcal{L}_{ce}$ can be computed by

$$\mathbb{E}_{D(\mathbf{p}|\boldsymbol{\alpha})}[\mathcal{L}_{ce}] = \int \left[\sum_{j=1}^{K} -y_j \log(p_j) \right] \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{K} p_j^{\alpha_j - 1} d\mathbf{p} = \sum_{j=1}^{K} y_j \left(\psi(S) - \psi(\alpha_j) \right), \tag{5}$$

where $\psi(\cdot)$ is the digamma function. By minimizing such risk, it is possible to ensure that correctly labeled observations generate as strong evidence as possible. Since the number of annotated pairs for ITM training is much larger than the number of categories for multi-classification, we simply regard "K" as the size of one training mini-batch, wherein visual and textual samples have a one-to-one correspondence. Therefore, such risk can be considered as the equivalent of the uncertainty-aware cross-entropy $\mathcal{L}_{uce}$ of TcVSE, which is defined as

$$\mathcal{L}_{uce}(\boldsymbol{\alpha}) = \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_{D(\mathbf{p}_i|\boldsymbol{\alpha}_i)} [\mathcal{L}_{ce}(\mathbb{I}_K, \mathbf{P}_i)], \tag{6}$$

where $\mathbb{I}_K$ is an identity matrix of size K. $\mathcal{L}_{uce}$ encourages VSE to generate as strong evidence as possible for positive pairs, which guarantees that evidence of positive pairs is higher than that of negative pairs. Furthermore, to further extreme the predicted evidence, we introduce Kullback-Leibler (KL) divergence to enforce the evidence of negative pairs to be zero. The penalization loss could be formulated as:

$$\begin{aligned}\mathcal{L}_{kl}(\boldsymbol{\alpha}) &= \frac{1}{K}\sum_{i=1}^{K}\mathbf{KL}\left[D\left(\mathbf{p}_i \mid \tilde{\boldsymbol{\alpha}}_i\right) \| D\left(\mathbf{p}_i \mid \langle 1,1,\cdots,1\rangle\right)\right] \\ &= \frac{1}{K}\sum_{k=1}^{K}\left[-\log\left(\Gamma(K)B\left(\tilde{\boldsymbol{\alpha}}_i\right)\right) + \sum_{k=1}^{K}\left(\tilde{\alpha}_{ik}-1\right)\left(\psi\left(\tilde{\alpha}_{ik}\right)-\psi\left(\tilde{S}_i\right)\right)\right],\end{aligned} \tag{7}$$

where $\tilde{S}_i = \sum_{j=1}^{K}\tilde{\alpha}_{ij}$, $\tilde{\boldsymbol{\alpha}}_i = \mathbb{I}_{K(i,:)} + \left(1-\mathbb{I}_{K(i,:)}\right)\odot\boldsymbol{\alpha}_i$, $\Gamma(\cdot)$ is the gamma function, and $\psi(\cdot)$ is the digamma function. Thus, the uncertainty-aware loss of one VSE branch (e.g., image-to-text) is given by

$$\mathcal{L}_u^{i2t} = \mathcal{L}_{uce}^{i2t} + \lambda\mathcal{L}_{kl}^{i2t}, \tag{8}$$

where λ is a balance factor that dynamically increases with the number of epochs. The dynamical strategy prevents the optimizer from overemphasizing the KL divergence at the beginning of training, otherwise, the optimizer will be misled by immature opinions leading to performance degradation. Finally, to simultaneously consider the bidirectional retrieval, we jointly optimize the two VSE branches as below:

$$\mathcal{L}_u = \mathcal{L}_u^{i2t} + \mathcal{L}_u^{t2i}, \tag{9}$$

where $\mathcal{L}_u^{t2i}$ is the evidential loss of *t2i* VSE branch, which could be computed like Equations (6) to (8).

2.3 Consistent Module

In our TcVSE, each branch focuses on different learning directions, due to the discrepancy between distinct retrieval tasks (one for image-to-text and another for text-to-image). Unfortunately, this will lead to various branches producing inconsistent uncertainty estimation, resulting in a performance drop. Specifically, given one query, one branch produces a prediction of low uncertainty, whereas the uncertainty of another branch might be higher as shown in Figure 2(b). Therefore, we introduce a consistency regularization to enforce the two VSE branches to produce consistent predictions on subjective opinions. To simplify presentation without losing generality, we only elaborate on the consistency loss of one direction (i.e., image-to-text) as follows:

$$\mathcal{L}_c^{i2t}\left(\mathbf{b}^{i2t}, \hat{\mathbf{b}}^{i2t}\right) = \frac{1}{K}\sum_{k=1}^{K}\left|b_k^{i2t} - \hat{b}_k^{i2t}\right|, \tag{10}$$

where $\mathbf{b}^{i2t}$ and $\hat{\mathbf{b}}^{i2t}$ are obtained from *i2t* and *t2i* branches with Equation (3), respectively. Similarly, we could easily obtain the consistency loss in another direction (e.g., text-to-image). Finally, the consistency loss $\mathcal{L}_c$ of our TcVSE could be formulated as:

$$\mathcal{L}_c = \frac{1}{K}\sum_{k=1}^{K}\left[\mathcal{L}_c^{i2t}\left(\mathbf{b}_k^{i2t}, \hat{\mathbf{b}}_k^{i2t}\right) + \mathcal{L}_c^{t2i}\left(\hat{\mathbf{b}}_k^{t2i}, \mathbf{b}_k^{t2i}\right)\right]. \tag{11}$$

The optimization process for our TcVSE is summarized in Algorithm 1.

3 Experiment

To evaluate our TcVSE, we conduct extensive experiments on two widely used benchmark datasets for Image-Text Matching. Following Lee et al. (2018), we measure the performance of image-to-text and text-to-image retrieval by Recall@K (K=1,5,10), which is defined as the proportion of correct items retrieved in the top K samples of the query. In addition, we adopt the sum of all Recall results to evaluate the overall performance.

Datasets	Flickr30K 1K Test							MS-COCO 5-fold 1K Test						
Visual Backbone: Faster-RCNN, Textual Backbone: Bi-GRU														
Eval Task	Image⟶Text			Text⟶Image				Image⟶Text			Text⟶Image			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
SCAN Lee et al. (2018)	67.4	90.3	95.8	48.6	77.7	85.2	465.0	72.7	94.8	98.4	58.8	88.4	94.8	507.9
CAMP Wang et al. (2019)	68.1	89.7	95.2	51.5	77.1	85.3	466.9	72.3	94.8	98.3	58.5	87.9	95.0	506.8
VSRN Li et al. (2019)	71.3	90.6	96.0	54.7	81.8	88.2	482.6	76.2	94.8	98.2	62.8	89.7	95.1	516.8
IMRAM Chen et al. (2020)	74.1	93.0	96.6	53.9	79.4	87.2	484.2	76.7	95.6	98.5	61.7	89.1	95.0	516.6
GSMN Liu et al. (2020)	76.4	94.3	97.3	57.4	82.3	89.0	496.7	78.4	96.4	98.6	63.3	90.1	95.7	522.5
SGRAF Diao et al. (2021)	77.8	94.1	97.4	58.5	83.0	88.8	499.6	79.6	96.2	98.5	63.2	90.7	96.1	524.3
VSE++* Chen et al. (2021)	62.2	86.6	92.3	45.7	73.6	81.9	442.3	68.5	92.6	97.1	54.0	85.6	92.7	490.5
VSE∞ Chen et al. (2021)	76.5	94.2	97.7	56.4	83.4	89.9	498.1	78.5	96.0	98.7	61.7	90.3	95.6	520.8
NCR Huang et al. (2021)	77.3	94.0	97.5	59.6	84.4	89.9	502.7	78.7	95.8	98.5	63.3	90.4	95.8	522.5
CGMN Cheng et al. (2022)	77.9	93.8	96.8	**59.9**	**85.1**	90.6	504.1	76.8	95.4	98.3	63.8	90.7	95.7	520.7
UARDA Zhang et al. (2022)	77.8	95.0	97.6	57.8	82.9	89.2	500.3	78.6	96.5	**98.9**	**63.9**	90.7	**96.2**	524.8
Our: TcVSE (i2t)	79.1	95.0	97.3	58.5	84.2	90.3	504.4	79.4	96.4	98.5	62.7	90.5	95.8	523.3
Our: TcVSE (t2i)	78.7	95.2	97.4	58.5	84.1	90.2	504.1	79.6	96.3	98.8	63.1	90.4	95.8	524.0
Our: TcVSE (i2t + t2i)	**80.3**	**95.5**	**97.8**	59.7	**85.1**	**90.9**	**509.3**	**80.6**	**96.7**	98.8	63.6	**90.8**	95.8	**526.3**
Visual Backbone: Faster-RCNN, Textual Backbone: Bert-base														
VSE++* Chen et al. (2021)	63.4	87.2	92.7	45.6	76.4	84.4	449.7	67.9	91.9	97.0	54.0	85.6	92.5	488.9
VSE∞ Chen et al. (2021)	81.7	95.4	97.6	61.4	85.9	91.5	513.5	79.7	96.4	98.9	64.8	91.4	96.3	527.5
VSRN++ Li et al. (2022a)	79.2	94.6	97.5	60.6	85.6	91.4	508.9	77.9	96.0	98.5	64.1	91.0	96.1	523.6
Our: TcVSE (i2t)	81.8	96.5	98.2	61.8	87.0	92.4	517.7	81.1	96.7	98.9	65.7	91.7	96.5	530.6
Our: TcVSE (t2i)	80.4	96.2	98.2	62.3	86.9	92.2	516.2	81.8	**97.1**	**99.0**	65.9	91.9	96.5	532.2
Our: TcVSE (i2t + t2i)	**82.9**	**97.0**	**98.5**	**63.9**	**88.3**	**93.1**	**523.7**	**82.2**	**97.1**	**99.0**	**66.7**	**92.2**	**96.7**	**533.9**

Table 1: Comparison of the bidirectional retrieval results (R@K %) on Flickr30K and MS-COCO datasets. i2t denotes that only the i2t VSE branch is used for evaluation. t2i means evaluate only with the t2i VSE branch. i2t+t2i reports the ensemble results of two branches. The best results are bolded in **black** and the second scores are in underline. For a convenience of analysis, the two backbone settings are abbreviated as "(Bi-GRU)" and "(Bert-base)".

3.1 Datasets and Implementation Details

Datasets: The benchmark datasets used in our experiments are Flickr30K Young et al. (2014) and MS-COCO Lin et al. (2014). Flickr30K is an image-text dataset collected from Flickr website and contains 31,000 images with five semantically correlated captions each. Following Lee et al. (2018), we adopt the same dataset splits in our experiments, i.e., 29,000 training images, 1,000 validation images, and 1,000 testing images. MS-COCO consists of 123,287 images, and each image also has five annotated text descriptions. Following Lee et al. (2018), 113,287 images for training, 5000 images for validation, and the remaining 5000 images for testing.

Implementation Detail. In our TcVSE, like VSE∞ Chen et al. (2021), a Faster-RCNN Anderson et al. (2018) detector (with ResNet-101) and Bi-GRU (or Bert-base Devlin et al. (2018)) serve as our visual and textual backbones, respectively. For each image, the visual backbone extracts the region proposals with top-36 confidence scores and projects each region into a 2,048-dimensional feature vector. Following Chen et al. (2021), we randomly discard some region proposals of each image to achieve augmentation during training. For each text description, we randomly mask some words to achieve data augmentation for each description. The dimensionality of the common embedding space is 1024. Different from most methods, we use the uncertainty-aware loss based on EDL for training, which additionally endows the model with the ability to uncertainty estimation.

We employ the AdamW optimizer Loshchilov & Hutter (2017) with weight decay factor 10e-4 to train the VSE branches. The learning rate of the visual model is 5e-4. For the textual model, the initial learning rate is 5e-4, except for Bert-base with 5e-5, and decaying by 10% every 10 epochs. The mini-batch size K is 128 with 25 training epochs on both Flickr30K and MS-COCO.

3.2 Comparison with State-of-the-art Methods

For a comprehensive evaluation, we compare our TcVSE with 12 state-of-the-art baselines, including SCAN Lee et al. (2018), CAMP Wang et al. (2019), VSRN Li et al. (2019), IMRAM Chen et al. (2020), GSMN Liu et al. (2020), SGRAF(SAF+SGR) Diao et al. (2021), VSE++* Chen et al. (2021), VSE∞ Chen et al. (2021), NCR Huang et al. (2021), CGMN Cheng et al. (2022), URDA Li et al. (2022a) and VSRN++ Li et al. (2022a). VSE++* is the basic version based on VSE∞ using *Average Pooling*. We conduct abundant comparison experiments as shown in Tables 1 and 2.

Furthermore, we also provide the comparison results compared with the state-of-the-art VSE-based methods in Table 5 for a comprehensive evaluation.

Results on Flickr30K. We report the experimental results on Flickr30K in Table 1. From the table, one could find that our TcVSE with a single VSE branch achieves comparable results, e.g., TcVSE (i2t) outperforms all baselines with rSum=504.4 under Bi-GRU and rSum=517.7 under Bert-base. Thanks to our trust-consistent learning, our TcVSE (i2t+t2i) is superior to all compared methods. Under the textual Bert-base backbone, our TcVSE could outperform all baselines with either one or two branches. Specifically, TcVSE (i2t+t2i) achieves remarkable improvement with the best R@1=82.9% for sentence retrieval and R@1=63.9% for image retrieval.

Results on MS-COCO. We present the qualitative results on MS-COCO with 5-fold 1K and full 5K test images in Tables 1 and 2, respectively. With Bi-GRU, our TcVSE could achieve a competitive performance compared to the state-of-the-arts. More specifically, TcVSE (i2t+t2i) achieves the best R@1 80.6% for sentence retrieval. In addition, Bert-base could further boost our TcVSE remarkably, i.e., a relative improvement of about 3% on R@1 compared to the best baseline VSE∞. In brief, our TcVSE with either one VSE branch or two branches could remarkably outperform all baselines, which demonstrates the effectiveness of our method.

For the experiments on MS-COCO 5K test images, the performance improvement is even more pronounced in terms of sentence retrieval, with a relative improvement of 7.7% (Bi-GRU) and 12.9% (Bert-base) on R@1 compared to best baselines. Both one and two branches of our TcVSE (Bert-base) could achieve conspicuous performance improvement. Furthermore, the consistent module could further boost the performance of TcVSE with one branch, which indicates that our trust-consistent learning will produce complementary and trustworthy predictions for retrieval improvement.

Datasets	MS-COCO 5K Test			
Faster R-CNN + Bi-GRU				
Eval Task	I$\longrightarrow$T		T$\longrightarrow$I	
Methods \ Metrics	R@1	R@10	R@1	R@10
SCAN (ECCV'18)	50.4	90.0	38.6	80.4
VSRN (ICCV'19)	53.0	89.4	40.5	81.1
IMARM (CVPR'20)	53.7	91.0	39.7	79.8
SGRAF (AAAI'21)	57.8	91.6	**41.9**	81.3
VSE++* (CVPR'21)	42.9	85.1	31.7	74.2
VSE∞ (CVPR'21)	56.6	91.4	39.3	81.1
NCR (NeurIPS'21)	58.2	91.5	41.7	81.3
CGMN (TOMM'22)	53.4	89.6	41.2	**82.4**
UARDA (TMM'22)	56.2	91.3	40.6	80.9
Our: TcVSE (i2t)	58.5	92.2	40.7	81.3
Our: TcVSE (t2i)	58.8	92.3	41.1	81.5
Our: TcVSE (i2t + t2i)	**60.5**	**92.5**	41.5	82.1
Faster-RCNN + Bert-base				
VSE++* (CVPR'21)	42.1	83.9	31.0	73.7
VSE∞ (CVPR'21)	56.6	91.4	39.3	81.1
VSRN++ (TPAMI'22)	54.7	90.9	42.0	82.7
Our: TcVSE (i2t)	60.7	93.2	43.6	83.7
Our: TcVSE (t2i)	61.4	93.4	43.6	84.0
Our: TcVSE (i2t + t2i)	**62.3**	**93.6**	**44.6**	**84.4**

Table 2: The evaluation results (R@K %) on MS-COCO 5K test set. More details of performance can be found in the Table 4.

3.3 Ablation Study

In this section, extensive ablation studies are carried out on Flickr30K to verify the contribution of each component to image-text matching. The experimental results are as shown in Table 3. We could comprehensively analyze the results from the following three distinct aspects:

Effectiveness. To verify the effectiveness of our EDL, we replace our evidential loss with *Max of Hinge Loss* (MH) Faghri et al. (2017) to optimize our VSE, i.e. #7 VSE with MH loss. From Table 3, one could see that other variants with EDL (i.e., #1–6) achieve better retrieval performance than VSE with MH loss, which indicates that our VSE endowed with EDL could remarkably improve performance by capturing the uncertainty. Moreover, our consistency module could further improve the retrieval performance of the two branches, even using only one branch for inference. More specifically, the module could relatively improve the performance by 1.65% (#1 vs. #2), 1.02% (#3 vs. #4), and 1.68% (#5 vs. #6), and in terms of R@1 for sentence retrieval, respectively. By fusing the two branches, our TcVSE could achieve further improvement, e.g., the full version of our TcVSE (#1) could relatively improve the version of one branch #3 and #5 by 1.52% and 2.03% in terms of R@1 for sentence retrieval, respectively.

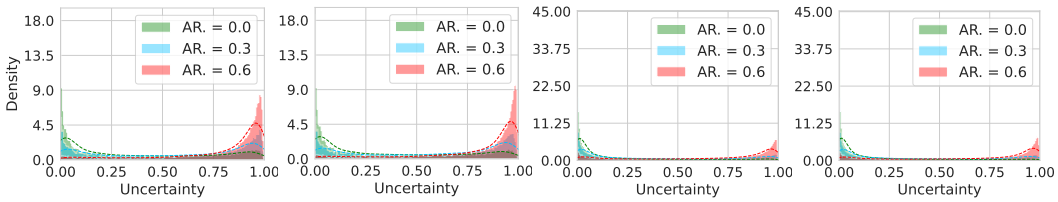
Complementarity. Two VSE branches are exploited to focus on different retrieval tasks, i.e., image-to-text and text-to-image matching. Obviously, such differences between tasks lead to distinct emphasis. Thus, aggregating the two VSE branches will take advantage of their complementary information, leading to further improvement, which has been verified by the results. Specifically, the variants with aggregation (i.e., #1 and #2) achieve better performance compared to the variants with single branches (i.e., #3-6).

Consistency. Thanks to our consistent module, the performance of our TcVSE could be improved even with only one single branch, i.e., #3 vs. #4, and #5 vs. #6. Hence, our consistent module could mutually promote the performance of different branches by eliminating the prediction discrepancy across different branches. Furthermore, the full version of TcVSE (#1) could achieve the best retrieval performance, which indicates that our consistent module not only mutually promotes the performance of each branch but also remains complementary information of different branches.

	Consistency	Branches		Image → Text			Text → Image			
No.	with $\mathcal{L}_c$	*i2t*	*t2i*	R@1	R@5	R@10	R@1	R@5	R@10	rSum
#1	✓	✓	✓	**80.3**	**95.5**	**97.8**	**59.7**	**85.1**	90.9	**509.3**
#2		✓	✓	79.0	94.9	97.8	59.0	84.9	**91.1**	506.7
#3	✓	✓		79.1	95.0	97.3	58.5	84.2	90.3	504.4
#4		✓		78.3	94.1	97.5	57.6	84.0	90.8	502.3
#5	✓		✓	78.7	95.2	97.4	58.5	84.1	90.2	504.1
#6			✓	77.4	94.5	97.7	58.0	83.6	90.1	501.3
#7	VSE with MH loss			75.7	93.5	97.3	56.3	82.4	89.3	494.5

Table 3: The impact of different TcVSE configurations on Flickr30K. For a convincing comparison, we report the results averaged over 3 replicates. The first column "Consistency" indicates whether consistency module is used to obtain consistent predictions. *i2t* and *t2i* denote the VSE branches used for performance evaluation, respectively.

3.4 Visualization of Uncertainty



(a) Flickr30K* (b) Flickr30K # (c) MS-COCO 5K* (d) MS-COCO 5K#

Figure 3: The visualization of the estimated uncertainty on Flickr30K and MS-COCO 5K test sets. The dashed lines are the curves fitted with Gaussian kernel functions. * means the sentence retrieval given image query and # expresses the image retrieval given sentence query.

To visually illustrate the uncertainty estimation, we plot the distribution diagrams of obtained uncertainty on the test sets of Flickr30K and MS-COCO. Since the intrinsic perturbation in data is uncontrollable and inconspicuous, it is hard to quantitatively evaluate the uncertainty estimated by the proposed method. To this end, we manually corrupt the inputs to amplify the unreliability of the data for easier observation, e.g., discard, swap, and mask operations used in Huang et al. (2021). Such data corruption could be seen as data augmentation. The proportion of corrupted image regions and words denotes the augmentation rate (AR). In the experiment, we investigate the uncertainty distribution quantified by our TcVSE (Bi-GRU) under three ARs (i.e., 0.0, 0.3, 0.6) as shown in Figure 3. From the figure, one could see that most retrievals under the low ARs have low certainty, i.e. clustering on the left. On the contrary, the uncertainty of the retrieval gradually increases as the ARs increase, as shown by most of the retrieval uncertainty gathered to the right in Figures 3(a) to 3(d). That is to say, as the AR increases, the correlation between image-text pairs will be degraded, resulting in increasing the retrieval uncertainty, which is consistent with the fact that data disturbance increases unreliability/uncertainty. Therefore, our method could effectively capture the uncertainty.

3.5 Qualitative Results

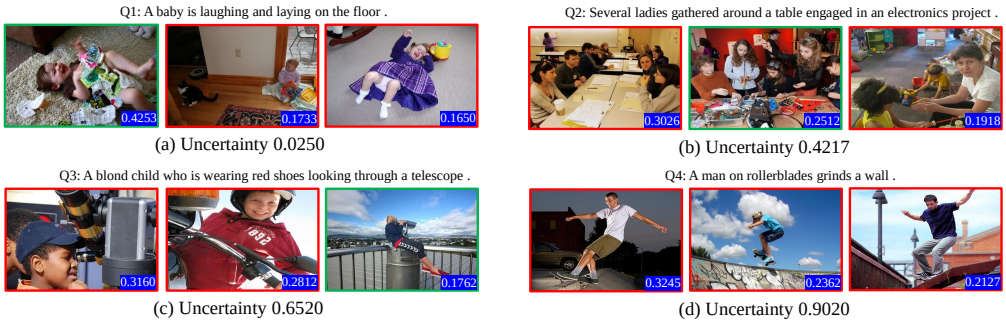
Figures 4 and 5 illustrate some qualitative cross-modal results retrieved by our TcVSE. In the figures, we also report the estimated uncertainty and the ensemble similarity measured by TcVSE for intuitive analysis. Unlike prior visual-textual matching methods, our TcVSE could quantify the overall uncertainty for cross-modal retrieval given each query, thus providing self-evaluation scores for the retrieved results. That is to say, our TcVSE could not only compute the similarities across different modalities for cross-modal retrieval inference but also self-evaluate the reliability of the results in terms of uncertainty, improving the interpretability of retrieval. For example, in Figure 4(a-c), the

predicted uncertainty by our TcVSE could self-evaluate the retrieval quality, namely more incorrect retrieved results with high uncertainty.

For example, in the completely correct examples (i.e. Figure 4(a) and Figure 5(a)), the correct retrieval is with high similarity and low overall uncertainty, which is viewed as trustworthy retrievals. In Figure 4(c) and Figure 5(d), retrieved results with high uncertainty are usually unreliable even with relatively high similarity, e.g., Figure 4(c) and Figure 5(d). More specifically, although the retrieved results have relatively high similarities compared to other correctly retrieved ones, they ignore/misunderstand some details in the queries, such as "one female" in Figure 4(c) and "skateboard" vs "rollerblade" in Figure 5(d). That is to say, it is very hard to evaluate the retrieval quality by the obtained similarities. Fortunately, our TcVSE could accurately estimate the uncertainty of the retrieved results leading to self-evaluating the retrieval quality.

(a) Uncertainty 0.0023 (b) Uncertainty 0.5951 (c) Uncertainty 0.8273

Figure 4: Some qualitative results of sentence retrieval on Flickr30K test set. We display the top-5 sentence retrievals ranked by inferred similarity of each image query. The correct sentence retrievals are marked with a green tick, e.g., (a.1), otherwise a red cross, e.g., (b.4). Meanwhile, we also give the estimated uncertainty and the ensemble similarity (bold font with bracket) of sentence retrievals in the sub-captions.



(a) Uncertainty 0.0250 (b) Uncertainty 0.4217

(c) Uncertainty 0.6520 (d) Uncertainty 0.9020

Figure 5: Some retrieved examples of top-3 image retrievals given a sentence on Flickr30K test set. We outline the correct retrievals in the green boxes and the incorrect ones in the red boxes. As with sentence retrieval, we give the estimated uncertainty and the ensemble similarity (white font with blue background) in the sub-captions.

4 Conclusion

In this paper, we revisit a practicable and meaningful problem in VSE-based image-text matching, i.e., "*How to make retrieval trustworthy?*". To this end, we present a Trust-consistent Visual Semantic Embedding method (TcVSE) for image-text matching, thus endowing the VSE models with the ability to self-evaluate the retrieval quality for trustworthy retrieval. Specifically, first, cross-modal evidential deep learning is proposed to capture accurate uncertainty of image-text matching. Second, a consistency module is presented to enforce the subjective opinion of distinct branches to be consistent for high reliability. Finally, we conduct extensive experiments and analyses to verify the effectiveness and self-evaluation of TcVSE.

REFERENCES

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13349–13358, 2021.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12655–12663, 2020.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15789–15798, 2021.

Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. Cross-modal graph matching network for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(4):1–23, 2022.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1218–1226, 2021.

Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *Advances in neural information processing systems*, 30, 2017.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Audun Jsang. Subjective logic: A formalism for reasoning under uncertainty. *Springer Verlag*, 2016.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

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.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11336–11344, 2020.

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.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022a.

Zheng Li, Caili Guo, Zerun Feng, Jenq-Neng Hwang, and Xijun Xue. Multi-view visual semantic embedding. IJCAI, 2022b.

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.

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10921–10930, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pp. 1–9, 2015.

Thomas P Minka. Bayesian inference, entropy, and the multinomial distribution. *Online tutorial*, 2003.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. 2020.

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.

Jonatas Wehrmann, Douglas M Souza, Mauricio A Lopes, and Rodrigo C Barros. Language-agnostic visual-semantic embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5804–5813, 2019.

Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 2088–2096, 2019.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Ronald R Yager and Liping Liu. *Classic works of the Dempster-Shafer theory of belief functions*, volume 219. Springer, 2008.

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.

Kun Zhang, Zhendong Mao, Anan Liu, and Yongdong Zhang. Unified adaptive relevance distinguishable attention network for image-text matching. *IEEE Transactions on Multimedia*, 2022.

Ke Zou, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. Tbrats: Trusted brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 503–513. Springer, 2022.

A APPENDIX

A.1 RELATED WORKS

Image-text matching. Most of the existing methods for image-text matching (ITM) could be roughly divided into two groups, i.e., the global-level methods represented by visual-semantic embedding (VSE) and the local-level methods with complex similarity inference. The global-level methods mainly aim to obtain good global representations from visual and textual modalities with the help of the well-designed feature extraction, enhancement, or aggregation strategy, and then directly compute the similarity, e.g., VSE++ Faghri et al. (2017), VSRN Li et al. (2019), and VSE∞ Chen et al. (2021). The local-level methods desire to learn the latent fine-grained alignments across different modalities for more accurate similarity inference, e.g., SCAN Lee et al. (2018), IMRAM Chen et al. (2020), SGRAF Diao et al. (2021), UARDA Zhang et al. (2022) and son on. Different from the mentioned lightweight methods, further breakthroughs have been made in the performance of downstream cross-modal tasks with the rapid development of large-scale visual language pre-training models in recent years, e.g., UNICODER-VL Li et al. (2020), CLIP Radford et al. (2021), and MaskCLIP Dong et al. (2022). However, the models are usually accompanied by high training or fine-tuning costs. In this paper, our research belongs to the lightweight global-level method.

Uncertainty-based learning. Deep learning has made promising progress in both academic research and industrial applications, but it is hard to quantify the uncertainty of deep models directly due to deterministic network prediction. Bayesian neural networks (BNNs) have been used to model uncertainty in computer vision tasks by placing priors over network deterministic weights, e.g., variational inference Kingma et al. (2015), approximations via dropout Gal & Ghahramani (2016); Gal et al. (2017), and so on. However, modeling uncertainty with BNNs is usually limited by the expensive sampling cost. Recently, Sensoy et al. (2018) proposed an uncertainty learning paradigm that combines evidence theory with DNNs, which places Dirichlet priors over discrete model predictions to directly model uncertainty with lower cost and it has been successfully applied in various tasks,

e.g., ClassificationSensoy et al. (2018); Han et al. (2022), RecognitionBao et al. (2021), and SegmentationZou et al. (2022). In this paper, we focus on the estimation of the uncertainty in image-text matching based on evidential deep learning.

A.2 DERIVATION

We carry out a detailed derivation process for some of the formulas in the paper.

The derivation of Equation (5):

$$\begin{aligned}\mathbb{E}_{D(\mathbf{p}|\boldsymbol{\alpha})}\left[\mathcal{L}_{ce}\right] &= \int\left[\sum_{j=1}^{K}-y_j\log\left(p_j\right)\right]\frac{1}{B\left(\boldsymbol{\alpha}\right)}\prod_{j=1}^{K}p_j^{\alpha_j-1}d\mathbf{p}\\ &= \sum_{j=1}^{K}y_j\left[\int\log\left(p_j\right)\frac{1}{B\left(\boldsymbol{\alpha}\right)}\prod_{j=1}^{K}p_j^{\alpha_j-1}d\mathbf{p}\right]\\ &= \sum_{j=1}^{K}y_j\mathbb{E}\left[\log\left(p_j\right)\right].\end{aligned}$$

From Minka (2003), $\mathbb{E}\left[\log\left(p_j\right)\right]$ could be as $\psi\left(S\right)-\psi\left(\alpha_j\right)$, where $S=\sum_{k=1}^{K}\alpha_k$. Thus,

$$\mathbb{E}_{D(\mathbf{p}|\boldsymbol{\alpha})}\left[\mathcal{L}_{ce}\right]=\sum_{j=1}^{K}y_j\left(\psi\left(S\right)-\psi\left(\alpha_j\right)\right).$$

The derivation of Equation (7):

$$\begin{aligned}\mathcal{L}_{kl}(\boldsymbol{\alpha}) &= \frac{1}{K}\sum_{i=1}^{K}\mathbf{KL}\left[D\left(\mathbf{p}_i\mid\tilde{\boldsymbol{\alpha}}_i\right)\|D\left(\mathbf{p}_i\mid\langle 1,1,\cdots,1\rangle\right)\right]\\ &= \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}\left[\log\frac{D\left(\boldsymbol{p}_i\mid\tilde{\boldsymbol{\alpha}}_i\right)}{D\left(\boldsymbol{p}_i\mid\langle 1,1,\cdots,1\rangle\right)}\right]\\ &= \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}\left[\log\left(\frac{\Gamma\left(\sum_{j=1}^{K}\tilde{\alpha}_{ij}\right)}{\Gamma(K)\prod_{j=1}^{K}\Gamma(\tilde{\alpha}_{ij})}\prod_{j=1}^{K}p_{ij}^{\tilde{\alpha}_{ij}-1}\right)\right]\\ &= \frac{1}{K}\sum_{i=1}^{K}\left(\log\left(\frac{\Gamma\left(\sum_{j=1}^{K}\tilde{\alpha}_{ij}\right)}{\Gamma(K)\prod_{j=1}^{K}\Gamma(\tilde{\alpha}_{ij})}\right)+\mathbb{E}\left[\log\prod_{j=1}^{K}p_{ij}^{\tilde{\alpha}_{ij}-1}\right]\right)\\ &= \frac{1}{K}\sum_{i=1}^{K}\left(-\log\left(\Gamma(K)B\left(\tilde{\boldsymbol{\alpha}}_i\right)\right)+\sum_{j=1}^{K}\left(\tilde{\alpha}_{ij}-1\right)\mathbb{E}\left[\log p_{ij}\right]\right)\\ &= \frac{1}{K}\sum_{i=1}^{K}\left[-\log\left(\Gamma(K)B\left(\tilde{\boldsymbol{\alpha}}_i\right)\right)+\sum_{k=1}^{K}\left(\tilde{\alpha}_{ik}-1\right)\left(\psi\left(\tilde{\alpha}_{ik}\right)-\psi\left(\tilde{S}_i\right)\right)\right]\end{aligned}$$

A.3 PSEUDOCODE

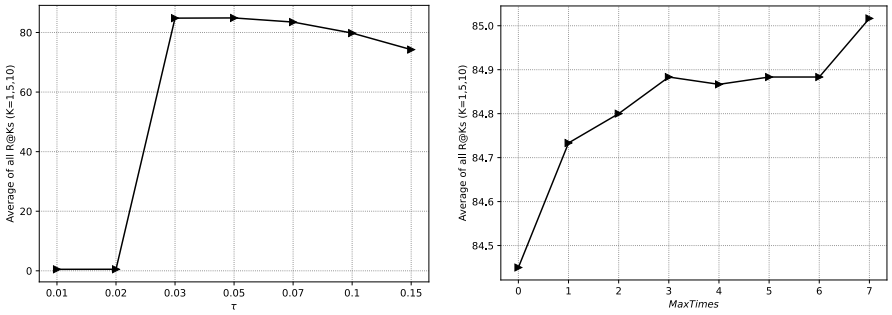
We provide the pseudocode of TcVSE (Algorithm 1) to help understand how TcVSE works.

A.4 PARAMETRIC ANALYSIS

TcVSE has two key hyper-parameters, i.e., τ in Equation (2) and $MaxTimes$ in Algorithm 1. Thus, we conduct detailed parameter experiments (shown in Figure 6) to evaluate the impact of different

Algorithm 1: *TcVSE*: Trust-consistent Visual Semantic Embedding pseudocode

Input: A well-paired subset $\{(\boldsymbol{u}_n, \boldsymbol{c}_n)\}_{n=1}^{N}$ of $(\mathcal{V}, \mathcal{C})$, temperature parameter τ.
Initialize: Initialize the parameters Θ of TcVSE.
while $e <$ *MaxEpoch* **do**
  for *x in Batches* **do**
    /*First step*/
    $x' = \text{Augment}(x)$
    $\{\boldsymbol{e}_k^{i2t}\}_{k=1}^{K} \longleftarrow \text{VSE}_{i2t}(x')$ \\ image-to-text
    $\{\boldsymbol{e}_k^{i2t}\}_{k=1}^{K} \longleftarrow \text{VSE}_{t2i}(x')$ \\ text-to-image
    for *each query* **do**
      Dirichlet distributions $D(\mathbf{p} \mid \boldsymbol{\alpha}) \longleftarrow \boldsymbol{e}$ \\ $\alpha = e + 1$
    end
    Obtain uncertainty-aware loss $\mathcal{L}_u$ with Equation (9)
    Θ = AdamW$(\mathcal{L}_u, \Theta)$
    /*Second step*/
    for $t < MaxTimes$ **do**
      Recompute $\{\boldsymbol{e}_k^{i2t}\}_{k=1}^{K}$ and $\{\hat{\boldsymbol{e}}_k^{i2t}\}_{k=1}^{K}$ \\ image-to-text
      for *each i2t query* **do**
        Obtain Subjective Opinions $\boldsymbol{b}^{i2t}, \hat{\boldsymbol{b}}^{i2t}$ with Equation (3)
      end
      Recompute $\{\hat{\boldsymbol{e}}_k^{t2i}\}_{k=1}^{K}$ and $\{\boldsymbol{e}_k^{t2i}\}_{k=1}^{K}$ \\ text-to-image
      for *each t2i query* **do**
        Obtain Subjective Opinions $\hat{\boldsymbol{b}}^{t2i}, \boldsymbol{b}^{t2i}$ with Equation (3)
      end
      Obtain the consistency loss $\mathcal{L}_c$ with Equation (11)
      Θ = AdamW$(\mathcal{L}_c, \Theta)$
    end
  end
end
Output: The learned parameters Θ



(a) Parametric analysis of τ (b) Parametric analysis of $MaxTimes$

Figure 6: We experiment with different settings of some hyper-parameters (i.e., τ and $MaxTimes$) of TcVSE for parametric analysis on Flickr30K. (a) is the visualization (the average of all R@ks (K=1, 5, 10)) of the parametric experiments for τ in Equation (2), and (b) is that of the parametric experiments for $MaxTimes$ in Algorithm 1.

hyper-parameter settings and obtain the better parameter settings for TcVSE. From Figure 6(a), TcVSE with too small τ will not be optimized well and perform poorly. Moreover, the performance of TcVSE gradually decreases from the best ($\tau = 0.03$) with the increment of τ, so we recommend setting τ for TcVSE within 0.03$\sim$0.05 to obtain stable and reliable performance. In all our experiments, τ is 0.05.

From Figure 6(b), as the $MaxTimes$ of consistency regularization increases, the better performance of TcVSE could be obtained due to the more consistent prediction, which is obviously reasonable. From the figures, one could find that when $MaxTimes$ is set to 3∼6, the performance gap is not large. In our experiments, we set $MaxTimes$ to 3.

A.5 SUPPLEMENTAL RESULTS

In this section, we supplement some experimental results. Specifically, we provide more detailed experimental results on the MS-COCO 5K test set and the results comparison of our TcVSE with the popular VSE-based methods (VSE++ Faghri et al. (2017), VSRN Li et al. (2019), LIWE Wehrmann et al. (2019), CVE Wang et al. (2020), VSE∞ Chen et al. (2021), VSRN++ Li et al. (2022a)),and MV-VSELi et al. (2022b). From Figure 5, our TcVSE achieves competitive results compared with that of the state-of-the-art image-text matching methods. Meanwhile, as shown in Table 5, compared with these popular VSE-based methods, TcVSE obviously achieves the best performance.

Datasets	MS-COCO 5K Test						
Image Backbone: Faster-RCNN, Text Backbone: Bi-GRU							
Eval Task	Image⟶Text			Image⟶Text			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	rSum
SCAN Lee et al. (2018)	50.4	82.2	90.0	38.6	69.3	80.4	410.9
CAMP Wang et al. (2019)	50.1	82.1	89.7	39.0	68.9	80.2	410.0
VSRN Li et al. (2019)	53.0	81.1	89.4	40.5	70.6	81.1	415.7
IMARM Chen et al. (2020)	53.7	83.2	91.0	39.7	69.1	79.8	416.5
SGRAF Diao et al. (2021)	57.8	-	91.6	**41.9**	-	81.3	-
VSE++* Chen et al. (2021)	42.9	74.5	85.1	31.7	61.8	74.2	370.2
VSE∞ Chen et al. (2021)	56.6	83.6	91.4	39.3	69.9	81.1	421.9
NCR Huang et al. (2021)	58.2	84.2	91.5	41.7	71.0	81.3	427.9
CGMN Cheng et al. (2022)	53.4	81.3	89.6	41.2	**71.9**	**82.4**	419.8
UARDA Zhang et al. (2022)	56.2	83.9	91.3	40.6	69.5	80.9	422.4
Our: TcVSE (i2t)	58.5	85.4	92.2	40.7	70.5	81.3	428.6
Our: TcVSE (t2i)	58.8	85.1	92.3	41.1	71.1	81.5	429.9
Our: TcVSE (i2t+t2i)	**60.5**	**86.0**	**92.5**	41.5	71.6	82.1	**434.2**
Image Backbone: Faster-RCNN, Text Backbone: Bert-base							
VSE++* Chen et al. (2021)	42.1	72.6	83.9	31.0	61.3	73.7	364.6
VSE∞ Chen et al. (2021)	56.6	83.6	91.4	39.3	69.9	81.1	421.9
VSRN++ Li et al. (2022a)	54.7	82.9	90.9	42.0	72.2	82.7	425.4
Our: TcVSE (i2t)	60.7	87.0	93.2	43.6	73.7	83.7	441.9
Our: TcVSE (t2i)	61.4	86.9	93.4	43.6	73.8	84.0	443.1
Our: TcVSE (i2t+t2i)	**62.3**	**87.9**	**93.6**	**44.6**	**74.6**	**84.4**	**447.4**

Table 4: The more detailed evaluation results (R@K %) on MS-COCO 5K test set

Datasets	Flickr30K 1K Test							MS-COCO 5-fold 1K Test						
Eval Task	Image⟶Text			Text⟶Image				Image⟶Text			Text⟶Image			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
VSE++Faghri et al. (2017)	52.9	80.5	87.2	39.6	70.1	79.5	409.8	64.6	90.0	95.7	52.0	84.3	92.0	479.6
LIWEWehrmann et al. (2019)	69.6	90.3	95.6	51.2	80.4	87.2	474.3	73.2	95.5	98.2	57.9	88.3	94.5	507.6
VSRN Li et al. (2019)	71.3	90.6	96.0	54.7	81.8	88.2	482.6	76.2	94.8	98.2	62.8	89.7	95.1	516.8
CVSE Wang et al. (2020)	73.6	90.4	94.4	56.1	83.2	90.0	487.7	78.6	95.0	97.5	66.3	91.8	96.3	525.5
VSE++′ Chen et al. (2021)	62.2	86.6	92.3	45.7	73.6	81.9	442.3	68.5	92.6	97.1	54.0	85.6	92.7	490.5
VSE∞′ Chen et al. (2021)	76.5	94.2	97.7	56.4	83.4	89.9	498.1	78.5	96.0	98.7	61.7	90.3	95.6	520.8
MV-VSE′Li et al. (2022b)	79.0	94.9	97.7	59.1	84.6	90.6	505.8	78.7	95.7	98.7	62.7	90.4	95.7	521.9
Our: TcVSE′	80.3	95.5	97.8	59.7	85.1	90.9	509.3	80.6	96.7	98.8	63.6	90.8	95.8	526.3
VSE++# Chen et al. (2021)	63.4	87.2	92.7	45.6	76.4	84.4	449.7	67.9	91.9	97.0	54.0	85.6	92.5	488.9
VSE∞# Chen et al. (2021)	81.7	95.4	97.6	61.4	85.9	91.5	513.5	79.7	96.4	98.9	64.8	91.4	96.3	527.5
VSRN++ Li et al. (2022a)	79.2	94.6	97.5	60.6	85.6	91.4	508.9	77.9	96.0	98.5	64.1	91.0	96.1	523.6
MV-VSE#Li et al. (2022b)	82.1	95.8	97.9	63.1	86.7	92.3	517.5	80.4	96.6	**90.0**	64.9	91.2	96.0	528.1
Our: TcVSE#	**82.9**	**97.0**	**98.5**	**63.9**	**88.3**	**93.1**	**523.7**	**82.2**	**97.1**	**99.0**	**66.7**	**92.2**	**96.7**	**533.9**

Table 5: Comparison with VSE-based methods on the bidirectional retrieval results (R@K %) on Flickr30K and MS-COCO datasets. ′ and # indicate that the textual backbones of VSE are Bi-GRU and Bert-base, respectively.

A.6 MORE RETRIEVAL RESULTS

1. A white dog is walking on top of a mountain . **(0.5655)** ✓
2. White dog on mountainside turns to face something offstage , sky in background . **(0.5450)** ✓
3. White dog standing on rock in front of blue sky **(0.4221)** ✓
4. A white dog is on a large rocky surface at a high elevation . **(0.3915)** ✓
5. A white dog is walking on a rocky hillside . **(0.3746)** ✓

(a) Uncertainty 0.0292

1. A group of asians smiling and eating . **(0.3128)** ✗
2. 6 people are gathered around to eat a big dinner . **(0.2562)** ✓
3. A family is sitting at a table ready to eat dinner . **(0.2560)** ✓
4. A family having a turkey dinner . **(0.2297)** ✓
5. Several people are eating lots of food . **(0.2092)** ✗

(b) Uncertainty 0.5765

1. A crowd of children in white shirts , black shorts and white hats jumping rope . **(0.2358)** ✓
2. A group of spectators watch a men 's sand volleyball game . **(0.1998)** ✗
3. A group of men in white shirts perform in a parade . **(0.1704)** ✗
4. A woman in a white shirt and hat speaks to a large crowd of men and women using a megaphone . **(0.1474)** ✗
5. Group of asian children dressed in white shirts and hats performing with a crowd looking on . **(0.1387)** ✓

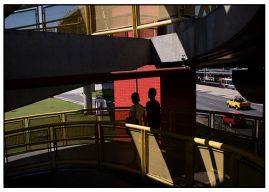
(c) Uncertainty 0.8277

1. A small boy enjoys chasing ducks in the fall . **(0.5607)** ✓
2. A child is enjoying the ducks on a fall day . **(0.5254)** ✓
3. A little boy chases ducks through the leaves . **(0.5065)** ✓
4. A smiling young boy plays in the leaves among the ducks . **(0.4901)** ✓
5. A little boy in a blue coat is playing next to some ducks in a park . **(0.4814)** ✓

(d) Uncertainty 0.0051

1. A smiling athlete in a black and red helmet holding ski poles . **(0.3591)** ✓
2. A man in a black , white , and red uniform is holding yellow ski poles and racing or preparing to race . **(0.2420)** ✓
3. A man in a black and white jersey is holding yellow ski poles and preparing to take off . **(0.2328)** ✓
4. A goalie looks down-field . **(0.2033)** ✗
5. hockey player in white uniform with stick **(0.1885)** ✗

(e) Uncertainty 0.5969



1. People stand outside behind a railing . **(0.2240)** ✗
2. Couple watching the moon on a bridge . **(0.1911)** ✗
3. Some young men looking over the edge of a bridge on a sunny day . **(0.1610)** ✗
4. People leaning on a railing . **(0.1466)** ✗
5. Several teenagers are looking over a railing in a dark room . **(0.1217)** ✗

(f) Uncertainty 0.9647

Figure 7: More qualitative results of sentence retrieval on Flickr30K test set. We display the top-5 sentence retrievals ranked by inferred similarity of each image query. The correct sentence retrievals are marked with a green tick, e.g., (a.1), otherwise a red cross, e.g., (b.1). Moreover, we also give the estimated uncertainty and the ensemble similarity (bold font with bracket) of sentence retrievals in the sub-captions. The conclusion is consistent with the analysis in the paper.

(a) Uncertainty 0.0214 (b) Uncertainty 0.0245

(c) Uncertainty 0.4231 (d) Uncertainty 0.4339

(e) Uncertainty 0.6489 (f) Uncertainty 0.9139

Figure 8: More retrieved examples of top-3 image retrievals given a sentence on Flickr30K test set. We outline the correct retrievals in the green boxes and the incorrect ones in the red boxes. As with sentence retrieval, we also give the estimated uncertainty and the ensemble similarity (white font with blue background) in the sub-captions. The conclusion is consistent with the analysis in the paper.