Supplementary Material: Deep Evidential Learning with Noisy Correspondence for Cross-modal Retrieval

Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, Peng Hu*

1 SUMMARY

In this supplementary material, we provide additional information. Specifically, we mainly supplement the detailed derivation of our evidential loss in Section 2.1. In Section 2.2, we elaborate the parameter settings for each dataset in the experiments. In Section 2.4, we describe the pseudocode of decl. In Section 2.4, we report experimental results on clean Flickr30K and MS-COCO without injecting manual noise. In Section 2.5, we conduct more experiments to analyze the contributions and impacts of the proposed loss functions. In Section 2.6, we report more qualitative retrieval results for insightful analysis.

2 CONTENT

2.1 The Derivations of Evidential Loss

In this section, we will deduce the evidential loss $\mathcal{L}_e$ of DECL in detail. $\mathcal{L}_e$ consists of two components, i.e., $\mathcal{L}_m$ (i.e., MSE or $l2$-norm loss in the form of evidence) and $\mathcal{L}_{kl}$. Given the parametrized Dirichlet distribution $\boldsymbol{\alpha}_i$ of a query I_i or T_i, the ground-truth $\boldsymbol{y_i}$, and density function $D\left(\boldsymbol{p}_i|\boldsymbol{\alpha}_i\right)$ mentioned in the paper, the more elaborated derivations could be shown below:

2.1.1 The Derivations of $\mathcal{L}_m$.

$$\begin{aligned}\mathcal{L}_m(\boldsymbol{\alpha}_i,\boldsymbol{y}_i) &= \int \left\|\boldsymbol{y}_i-\boldsymbol{p}_i\right\|_2^2 \frac{1}{B\left(\boldsymbol{\alpha}_i\right)}\prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} d\boldsymbol{p}_i \\ &= \sum_{j=1}^{K}\mathbb{E}\left[y_{ij}^2-2y_{ij}p_{ij}+p_{ij}^2\right] \\ &= \sum_{j=1}^{K}\left(y_{ij}^2-2y_{ij}\mathbb{E}(p_{ij})+\mathbb{E}\left(p_{ij}^2\right)\right) \\ &= \sum_{j=1}^{K}\left[\left(y_{ij}-\mathbb{E}(p_{ij})\right)^2+\mathrm{Var}\left(p_{ij}\right)\right] \\ &= \sum_{j=1}^{K}\left(y_{ij}-\frac{\alpha_{ij}}{L_i}\right)^2+\frac{\alpha_{ij}\left(L_i-\alpha_{ij}\right)}{L_i^2\left(L_i+1\right)} \\ &= \sum_{j=1}^{K}\left[\left(y_{ij}-\hat{p}_{ij}\right)^2+\frac{\hat{p}_{ij}\left(1-\hat{p}_{ij}\right)}{\left(L_i+1\right)}\right],\end{aligned} \tag{1}$$

where $L_i=\sum_{j=1}^{K}\alpha_{ij}$, and $\mathbb{E}(p_{ij})$ and $\mathrm{Var}\left(p_{ij}\right)$ are the expected value and the variance of p_{ij}, respectively. Following [6], $\mathbb{E}(p_{ij})$ could be obtained by

$$\begin{aligned}\mathbb{E}(p_{ij}) &= \int\cdots\int p_{ij}\, D\left(\boldsymbol{p}_i|\boldsymbol{\alpha}_i\right) dp_{i1}\ldots dp_{iK} \\ &= \int\cdots\int p_{ij}\,\frac{\Gamma\left(L_i\right)}{\prod_{j=1}^{K}\Gamma(\alpha_{ij})}\prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} dp_{i1}\ldots dp_{iK}\end{aligned} \tag{2}$$

Move $p_{ij}^{\alpha_{ij}-1}$ out of the product:

$$\mathbb{E}(p_{ij}) = \int\cdots\int p_{ij}^{\alpha_{ij}}\,\frac{\Gamma\left(L_i\right)}{\prod_{j=1}^{K}\Gamma(\alpha_{ij})}\prod_{j'\neq j} p_{ij'}^{\alpha_{ij'}-1} dp_{i1}\ldots dp_{iK}$$

According to the Gamma function's property, i.e., $\Gamma\left(x+1\right)=x\Gamma\left(x\right)$, $\mathbb{E}(p_{ij})$ can be written as follows:

$$\mathbb{E}(p_{ij}) = \int\cdots\int \frac{\alpha_{ij}}{L_i}\,\frac{\Gamma\left(L_i+1\right)}{\prod_{j=1}^{K}\Gamma\left(\beta_j\right)}\prod_{j=1}^{K} p_{ij}^{\beta_j-1} dp_{i1}\ldots dp_{iK}=\frac{\alpha_{ij}}{L_i},$$

where $\{\beta_j\}_{j=1}^{K}$ is a new set of Dirichlet parameters with $\beta_j=\alpha_{ij}+1$. Because the Dirichlet with new parameters β_j must nevertheless integrate to 1, we could obtain the final result $\mathbb{E}(p_{ij})=\frac{\alpha_{ij}}{L_i}$, which could be regarded as an estimate of the query probability p_{ij}, i.e., $\hat{p}_{ij}$ [7].

2.1.2 The Derivations of $\mathcal{L}_{kl}$. According to the definition in the paper, two new density functions for calculating Kullback-Leibler (KL) divergence are

$$D\left(\boldsymbol{p}_i \mid \tilde{\boldsymbol{\alpha}}_i\right)=\frac{\Gamma\left(\sum_{j=1}^{K}\tilde{\alpha}_{ij}\right)}{\prod_{j=1}^{K}\Gamma(\tilde{\alpha}_{ij})}\prod_{j=1}^{K} p_{ij}^{\tilde{\alpha}_{ij}-1} \tag{3}$$

and

$$D\left(\boldsymbol{p}_i \mid \mathbf{1}\right)=\Gamma(K). \tag{4}$$

Therefore, our KL divergence loss could be obtained by

$$\begin{aligned}&\mathcal{L}_{kl}(\boldsymbol{\alpha}_i,\boldsymbol{y}_i) \\ &= KL\left[D\left(\boldsymbol{p}_i \mid \tilde{\boldsymbol{\alpha}}_i\right)\|D\left(\boldsymbol{p}_i \mid \mathbf{1}\right)\right] \\ &= \mathbb{E}\left(\log\frac{D\left(\boldsymbol{p}_i \mid \tilde{\boldsymbol{\alpha}}_i\right)}{D\left(\boldsymbol{p}_i \mid \mathbf{1}\right)}\right)=\mathbb{E}\left(\log\left(\frac{\Gamma\left(\sum_{j=1}^{K}\tilde{\alpha}_{ij}\right)}{\Gamma(K)\prod_{j=1}^{K}\Gamma(\tilde{\alpha}_{ij})}\prod_{j=1}^{K} p_{ij}^{\tilde{\alpha}_{ij}-1}\right)\right) \\ &= \log\left(\frac{\Gamma\left(\sum_{j=1}^{K}\tilde{\alpha}_{ij}\right)}{\Gamma(K)\prod_{j=1}^{K}\Gamma(\tilde{\alpha}_{ij})}\right)+\mathbb{E}\left(\log\prod_{j=1}^{K} p_{ij}^{\tilde{\alpha}_{ij}-1}\right) \\ &= \log\left(\frac{\Gamma\left(\sum_{j=1}^{K}\tilde{\alpha}_{ij}\right)}{\Gamma(K)\prod_{j=1}^{K}\Gamma\left(\tilde{\alpha}_{ij}\right)}\right)+\sum_{j=1}^{K}\left(\tilde{\alpha}_{ij}-1\right)\mathbb{E}(\log p_{ij}), \\ &= \log\left(\frac{\Gamma\left(\sum_{j=1}^{K}\tilde{\alpha}_{ij}\right)}{\Gamma(K)\prod_{j=1}^{K}\Gamma\left(\tilde{\alpha}_{ij}\right)}\right)+\sum_{j=1}^{K}\left(\tilde{\alpha}_{ij}-1\right)\left[\psi\left(\tilde{\alpha}_{ij}\right)-\psi\left(\sum_{j=1}^{K}\tilde{\alpha}_{ij}\right)\right],\end{aligned} \tag{5}$$

For the expection $\mathbb{E}(\log p_{ij})$, [6] provides a solution, i.e.,

$$\begin{aligned}\mathbb{E}(\log p_{ij}) &= \frac{\partial}{\partial\alpha_{ij}}\sum_{j=1}^{K}\log\Gamma\left(\tilde{\alpha}_{ij}\right)-\log\Gamma\left(\sum_{j=1}^{K}\tilde{\alpha}_{ij}\right) \\ &= \psi\left(\tilde{\alpha}_{ij}\right)-\psi\left(\sum_{j=1}^{K}\tilde{\alpha}_{ij}\right),\end{aligned} \tag{6}$$

*Corresponding author: Peng Hu (penghu.ml@gmail.com).

where $\tilde{\boldsymbol{\alpha}}_i = \boldsymbol{y}_i + (1 - \boldsymbol{y}_i) \odot \boldsymbol{\alpha}_i$ is the Dirichlet parameters after removing the unreliable evidence from $\boldsymbol{\alpha}_i$, and $\Gamma(\cdot)$ and $\psi(\cdot)$ are the gamma and digamma functions, respectively.

2.2 Parameter Settings

See Table 1 for some specific experimental parameters. Besides, we select the checkpoint with the best performance on the validation set for testing.

2.3 Experimental Results Without Noise

We conduct comparison experiments in terms of cross-modal retrieval on two datasets to evaluate the performance of our DECL without simulated noise. The baselines are SCAN [4], VSRN [5], IMRAM [1], SGRAF [2] and, NCR [3], respectively. Our DECL achieves competitive results. In Flickr30K, DECL remarkably improves the performance of baselines SAF and SGR, respectively, which shows effectiveness in the absence of simulated noise. Moreover, DECL-SGRAF achieves the best overall performance (**505.0**) compared to all baselines. In MS-COCO, although DECL-SGRAF does not achieve the best overall performance, the results are still competitive, with the best R@5=**96.3%** and R@10=**63.3%** for image-to-text retrieval.

2.4 Pseudocode of DECL

See Algorithm 1 for the pseudo code of DECL.

Algorithm 1: Deep Evidential Cross-Modal Learning

Input: Given the annotated cross-modal pairs $\mathcal{P} = \{P_i\}_{i=1}^{N}$, $P_i = (I_i, T_i)$, the cross-modal model $\mathcal{M}$, and $Step = 0$.

Warmup the model $\mathcal{M}$ using $\mathcal{L}_{overall}$ with mini-batch manner ;
for $e = 1 : epochs$ **do**
  Predict all alignment labels $\{l_i\}_{i=1}^{N}$ for all pairs;
  for $b = 1 : batchs$ **do**
    Update $Step = Step + 1$;
    Update the dynamic parameter n with $Step$;
    for $i = 1 : K$ **do**
      Get the query evidence vectors $\boldsymbol{e}_i^{i2t}, \boldsymbol{e}_i^{i2t} \in \mathbb{R}^K$ of P_i, respectively.
      Get the corresponding Dirichlet distributions $\boldsymbol{\alpha}_i^{i2t}, \boldsymbol{\alpha}_i^{t2i} \in \mathbb{R}^K$ according to $\boldsymbol{e}_i^{i2t}$ and $\boldsymbol{e}_i^{i2t}$, respectively.
      Get the vector $\boldsymbol{y}_i$ according alignment label l_i.
      Compute bidirectional evidential loss $\mathcal{L}_e$ of P_i.
      Compute RDH loss $\mathcal{L}_h$ of P_i.
    end
    Compute the overall loss $\mathcal{L}_{overall}$ of mini-batch to conduct positive and negative learning with Adam optimizer.
  end
end

Output: Model $\mathcal{M}$

2.5 Evidence analysis

In our DECL, we extend the Dirichlet distribution form of the least squares loss (SL), namely $\mathcal{L}_m$, which could be called MSE or l_2-norm loss in the form of evidence. To analyze its effectiveness, we conduct a comprehensive comparison experiment. Specifically, we use the general MSE loss and $\mathcal{L}_m$ to replace the loss function of the original SGR for training and report the retrieval results on Flickr30K with different noise rates (i.e., 20%, 40%, 60%, and 80%) as shown in Table 3. From the experimental results, one could find that although the partial results of $\mathcal{L}_m$ at 40% and 60% noise are competitive, the best performance is achieved at all noise rate settings.

2.6 Qualitative Results

In the section, we present more qualitative results in Figures 1 and 2. From the retrieved results, one could see that our DECL could capture the semantics of images or sentences. Specifically, regardless of whether the correctly retrieved or incorrectly retrieved samples, one could find that they have some common semantic correlation with the given query. For example, in Figure 1(b), the incorrect retrieved result also captures the semantic properties of the image, i.e., "a man", "in a hat", and"in glassess". Furthermore, the uncertainty quantified by DECL reflects the quality of cross-modal retrieval, i.e., confidence. Specifically, the results with low uncertainty (e.g., Figure 1(a) and Figure 2(a)) are confidently correct, while the results with high uncertainty (e.g., Figure 1(e,f) and Figure 2(e,f)) are likely incorrect. However, similarity has no deterministic relationship with uncertainty, i.e., the results with higher similarity maybe have higher overall uncertainty, e.g., Figures 1(e)-1,2,3 and (f)-1,2 in image-to-text retrieval. Thus, it is hard for the traditional cross-modal models to directly estimate the uncertainty/confidence without a specific uncertainty learning paradigm like our DECL.

REFERENCES

[1] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. 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*. 12655–12663.

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

[3] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. 2021. Learning with Noisy Correspondence for Cross-modal Matching. *Advances in Neural Information Processing Systems* 34 (2021), 29406–29419.

[4] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*. 201–216.

[5] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4654–4662.

[6] Thomas Minka. 2000. Estimating a Dirichlet distribution.

[7] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems* 31 (2018).

		Training parameters					Model parameters				
Noise	Dataset	warmup_epochs	epochs	lr_update	batch size (K)	learning_rate	λ_1	λ_2	η	μ	τ
0%	Flickr30K	5	40	20	128	0.0002	0.8	0.1	0.025	5	0.1
	MS-COCO	2	20	10	128	0.0005	0.8	0.1	0.025	5	0.1
	CC-152K	2	20	10	128	0.0002	0.8	0.1	0.025	5	0.1
20%,40%	Flickr30K	5	40	20	128	0.0002	0.8	0.1	0.025	5	0.1
	MS-COCO	2	20	10	128	0.0005	0.8	0.1	0.025	5	0.1
60%,80%	Flickr30K	2	40	20	128	0.0002	0.8	0.1	0.025	5	0.1
	MS-COCO	1	20	10	128	0.0005	0.8	0.1	0.025	5	0.1

Table 1: The settings of some key parameters for training on three datasets. warmup_epochs means the epochs for warmuping model and DECL decays the leaning rate (lr) by 0.1 in lr_update epoch. λ_1 and λ_2 are trade-off parameters of DECL. η and μ are the annealing coefficient and lower bound of our Robust Dynamic Hinge loss $\mathcal{L}_h$. τ is the scaling parameter of the evidence extractor.

		Flickr30K							MS-COCO 1K						
		Image → Text			Text → Image				Image → Text			Text → Image			
Noise	Methods	R@1	R@5	R@10	R@1	R@5	R@10	Sum	R@1	R@5	R@10	R@1	R@5	R@10	Sum
0%	SCAN [4]	67.4	90.3	95.8	48.6	77.7	85.2	465.0	69.2	93.6	97.6	56.0	86.5	93.5	496.4
	VSRN [5]	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 [1]	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
	SAF [2]	73.7	93.3	96.3	56.1	81.5	88.0	488.9	76.1	95.4	98.3	61.8	89.4	95.3	516.3
	SGR [2]	75.2	93.3	96.6	56.2	81.0	86.5	488.9	78.0	95.8	98.2	61.4	89.3	95.4	518.1
	SGRAF [2]	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**
	NCR [3]	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
	NCR* [3]	72.7	91.8	95.8	55.7	82.3	88.3	486.6	75.9	95.4	98.0	61.1	89.2	95.1	514.7
	DECL-SAF	77.0	93.9	96.9	56.8	81.7	88.0	494.3	77.8	95.8	98.4	61.4	89.2	95.2	517.8
	DECL-SGR	77.1	93.6	96.7	57.3	82.1	88.4	495.2	76.9	95.8	98.6	61.6	89.4	95.2	517.5
	DECL-SGRAF	**79.8**	**94.9**	97.4	59.5	83.9	89.5	**505.0**	79.1	**96.3**	**98.7**	**63.3**	90.1	95.6	523.1

Table 2: Performance comparison without simulated noisy correspondence (0% noise) on Flickr30K and MS-COCO 1K. NCR* means that a single branch of NCR is used for validation instead of averaging the similarity calculated by the two branches for validation.

		Image → Text			Text → Image			
Noise	Loss	R@1	R@5	R@10	R@1	R@5	R@10	Sum
20%	MSE	71.6	91.6	96.0	52.1	77.8	84.9	474.0
	MSE($\boldsymbol{\alpha}$)	**72.3**	**93.5**	**96.7**	**53.5**	**79.0**	**86.0**	**481.0**
40%	MSE	67.9	89.2	93.8	**49.1**	**74.5**	82.0	456.5
	MSE($\boldsymbol{\alpha}_i$)	**68.0**	**89.5**	**94.1**	48.7	**74.5**	**82.1**	**456.9**
60%	MSE	61.1	85.2	**91.6**	42.2	68.5	**77.0**	425.6
	MSE($\boldsymbol{\alpha}_i$)	**62.2**	**86.0**	91.2	**42.7**	**69.0**	76.4	**427.5**
80%	MSE	40.6	66.9	75.1	26.9	49.8	**60.5**	319.8
	MSE($\boldsymbol{\alpha}_i$)	**44.3**	**68.4**	**78.3**	**29.2**	**50.7**	58.9	**329.8**

Table 3: MSE($\boldsymbol{\alpha}_i$) loss is $\mathcal{L}_m\left(\boldsymbol{\alpha}_i, \boldsymbol{y}_i\right)$ defined by Equation 1. MSE is commonly defined as $\frac{1}{K}\left\|\boldsymbol{y}_i - \boldsymbol{s}_i\right\|_2^2$, where $\boldsymbol{s}_i \in \mathbb{R}^K$ is the query similarity vector of query I_i or T_i. Other configurations are the same as DECL.

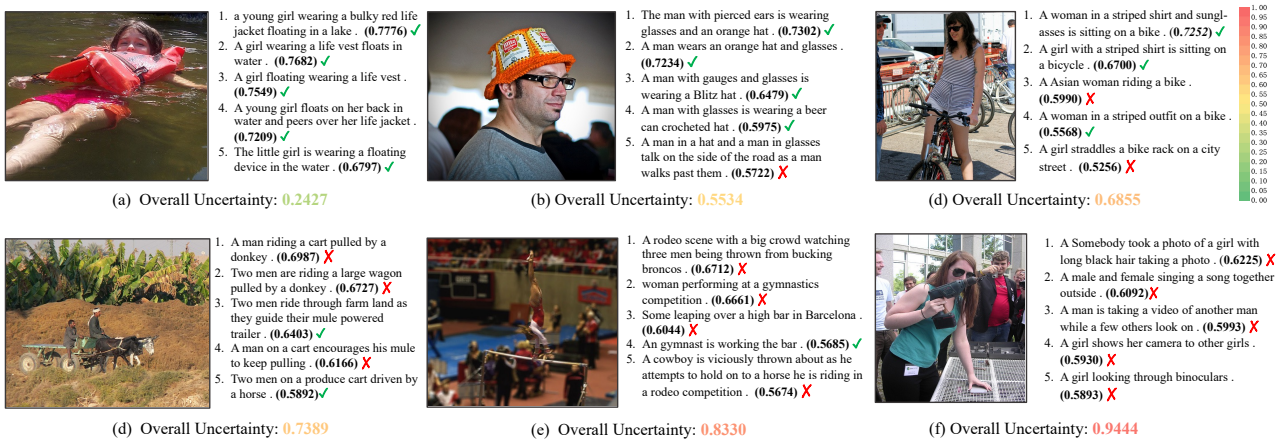

Figure 1: Some retrieved examples of cross-modal retrieval on Flickr30K. For each image query, we show the top-5 ranked sentences. The correctly matched marked are with the green check marks, otherwise the red cross. Estimated uncertainty and the original similarity (bold font in bracket) are given for all retrieved results.

Figure 2: Some retrieved examples of cross-modal retrieval on Flickr30K. For each sentence query, we show the top-3 ranked images, ranking from left to right. We outline the correctly matched images in green boxes and mismatched in red boxes. Estimated uncertainty and the original similarity (white font with blue background) are given for all retrieval.