

Source-free Image-text Matching via Uncertainty-aware Learning

Mengxiao Tian^{1,2}, Shuo Yang^{2*}, Xinxiao Wu^{1,2}, *Member, IEEE*, Yunde Jia², *Member, IEEE*,

Abstract—When applying a trained image-text matching model to a new scenario, the performance may largely degrade due to domain shift, which makes it impractical in real-world applications. In this paper, we make the first attempt on adapting the image-text matching model well-trained on a labeled source domain to an unlabeled target domain in the absence of source data, namely, source-free image-text matching. This task is challenging since it has no direct access to the source data when learning to reduce the domain shift. To address this challenge, we propose a simple yet effective method that introduces uncertainty-aware learning to generate high-quality pseudo-pairs of image and text for target adaptation. Specifically, starting with using the pre-trained source model to retrieve several top-ranked image-text pairs from the target domain as pseudo-pairs, we then model uncertainty of each pseudo-pair by calculating the variance of retrieved texts (resp. images) given the paired image (resp. text) as query, and finally incorporate the uncertainty into an objective function to down-weight noisy pseudo-pairs for better training, thereby enhancing adaptation. This uncertainty-aware training approach can be generally applied on all existing models. Extensive experiments on the COCO and Flickr30K datasets demonstrate the effectiveness of the proposed method.

Index Terms—Source-free adaptation, uncertainty-aware learning, image-text matching

I. INTRODUCTION

IMAGE-text matching has achieved remarkable progress in a variety of applications, such as cross-modal retrieval [1], [2], [3], [4], [5], [6], image captioning [7], [8], [9], [10], [11], and visual question answering [12], [13], [14], [15], [16]. Existing methods of image-text assume that the training data and test data all come from the same distribution, which rarely holds true in real-world applications. Directly applying a well-trained image-text matching model to a new scenario (target domain) lying in different data distribution from the training instances (source domain) easily suffers from severe performance degradation. Such distribution discrepancy termed as domain shift results from both visual and language variations. Accessing both labeled source and unlabeled target data can better characterize the domain shift, however, it is a major bottleneck for real-world deployment scenarios, due to the source data privacy and limited memory storage in small devices. To that end, we propose a source-free adaptation framework of image-text matching, where only the pre-trained

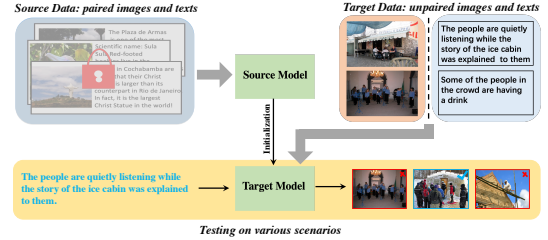


Fig. 1: Illustration of the source-free image-text matching framework. Given an image-text matching model well-trained on the source data and unlabeled target data, the goal of source-free image-text matching is to adapt the source model to the target domain.

source model and the unlabeled target data are available during training, as illustrated in Fig. 1.

Under this framework, we propose an uncertainty-aware learning method, which generates reliable pseudo image-text pairs to achieve good adaptation of the source model to the target domain. Specifically, we use an image-text matching model well-trained on the source domain to retrieve the most matched image-text pairs from the target domain, called pseudo-pairs of image and text, as training samples to fine-tune the source model for adaptation. However, due to the domain shift caused by applying a trained image-text matching model to a new scenario, these pseudo-pairs are noisy, referring to alignment errors present in multimodal data. Some relevant methods [17], [18] have also explored such a similar problem, but with different motivations and methodologies. Under such noisy data conditions, the performance of adaptation may witness a substantial drop. In some related approaches [19], [20], [21], [22], variance has been widely adopted as a common method for quantifying uncertainty, owing to its simplicity and popularity in statistical science. These approaches investigate the relationship between uncertainty and noisy data, with researchers considering uncertainty as an indicator of the quality of noisy data. Due to the low quality of noisy data, noisy labels could bring more uncertainty. Inspired by that, we introduce uncertainty information to measure the noise of each pseudo-pair, where the higher the uncertainty is, the more serious the noise is. The uncertainty of each pseudo-pair is represented by the variance of the retrieved texts (resp. images) from the target domain given the pseudo-paired image (resp. text). Finally, we incorporate the uncertainty into training objective to penalize the pseudo-pairs with high uncertainty to update the model, and thus down-weight noisy pseudo-pairs to enhance the model adaptation to the target. It is worth noting that our method is

Manuscript received 9 July 2024; revised 26 September 2024; accepted 22 October 2023. This work was supported in part by the Natural Science Foundation of China (NSFC) under Grants. No. 62072041 and No. 62176021.

¹Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China.

²Guangdong Provincial Lab of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China.

*Corresponding author. E-mail: yangshuo@smbu.edu.cn

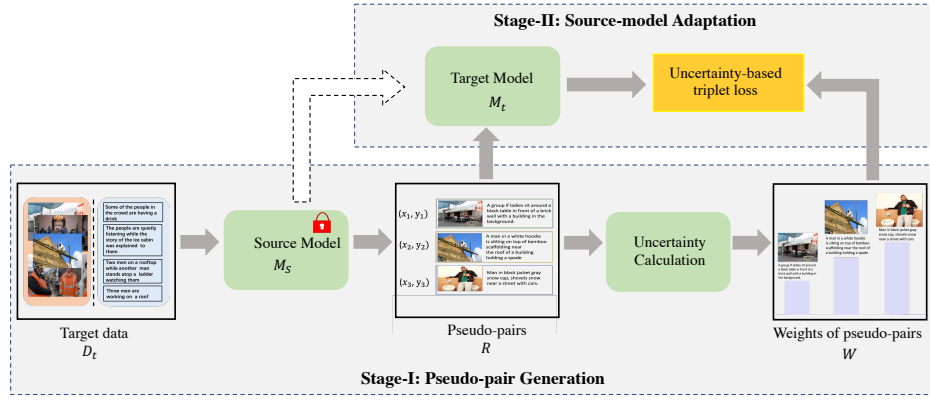


Fig. 2: Overview of the proposed source-free image-text matching method by using uncertainty-aware learning.

a plug-and-play module compatible with all existing image-text matching models, due to the fact that our pseudo-paired uncertainty can be calculated offline, and our method enhances the adaptability of these models to new scenarios with different data distributions.

In summary, our contributions are: (1) We introduce a source-free adaptation framework for image-text matching to adapt a source model well to a target domain without source data. (2) We propose an effective uncertainty-aware learning method for generating high-quality pseudo-pairs to facilitate target adaptation. (3) Extensive experiments on the COCO and Flickr30K datasets demonstrate the effectiveness of the proposed method with significant improvements.

II. THE PROPOSED METHOD

A. Overview

We propose a source-free adaptation framework for image-text matching, allowing the adaptation of a source model to a target domain without accessing the source data. Our method includes two stages: pseudo-pair generation and source-model adaptation. In the first stage, we retrieve top- k texts/images with the highest similarity scores from the target domain, based on an image-text matching model trained on the source domain. We then measure each pseudo-pair's uncertainty by calculating the variance of retrieved samples. In the second stage, we fine-tune the source model using these pseudo-pairs, incorporating their uncertainties into the training loss to penalize noisy pseudo-pairs with high uncertainty, thus ensuring effective adaptation. Fig. 2 illustrates the overview of our method.

B. Pseudo-pair Generation

Image-to-Text Retrieval. We apply the image-text matching model trained on the source domain to the target domain to generate pseudo-pairs of image and text. Specifically, for a given image x_i from the target domain, we feed it into the image-specific network g of the given source model M_s to generate the visual embedding feature $g(x_i)$. Similarly, for a language description y_i from the target domain, we employ the text-specific network f of the given source model to generate the textual embedding feature $f(y_i)$. We retrieve the top- K texts (resp. images) from all the candidate texts (resp. images) by ranking their similarity scores to a specific target image (resp.

text), and then select the top-ranking pair to form the initial pseudo-pairs for target domain, which are used for adapting the source model.

Pseudo-pair Uncertainty. The generated pseudo-pairs contain noise due to domain shift, which can mislead source model adaptation and degrade performance on the target domain. To mitigate this, we introduce uncertainty as a measure of noise for each pseudo-pair. Higher uncertainty values indicate lower reliability and greater noise. We model the uncertainty of each pseudo-pair by calculating the variance of the retrieved texts (resp. images) given the pseudo-paired image (resp. text) as query. Specifically, the weighted average textual embedding features for the top- K ($K = 300$) retrieved texts given a query image are calculated as

$$\mu^T = \frac{\sum_{i=1}^K w_i f(y_i^T)}{\sum_{i=1}^K w_i}, \quad (1)$$

where w_i denotes the initial weight of each retrieved candidate text. The weights are assigned based on candidate rankings using $w_i = 1 - \frac{r_i - 1}{K - 1}$, where r_i is the rank (with $r_i = 1$ for the highest-ranked candidate), ensuring all weights are within $[0, 1]$. In this case, a higher initial weight is assigned to a retrieved text with high-ranking score, while a lower initial weight is assigned to one with low-ranking score. This weighting strategy stems from the notion that the retrieved texts with higher ranking scores are generally greater relevant than their lower-ranking counterparts.

Symmetrically, the weighted average embedding visual features μ^I for the top- K retrieved images given a query text can be formulated by interchanging $f(y_i^T)$ and $g(x_i^I)$ in Eq. 1. Utilizing these dual-weighted average embedding features, their respective variances can be determined as

$$V^T = \frac{\sum_{i=1}^K w_i (f(y_i^T) - \mu^T)^2}{\sum_{i=1}^K w_i}, V^I = \frac{\sum_{i=1}^K w_i (g(x_i^I) - \mu^I)^2}{\sum_{i=1}^K w_i}, \quad (2)$$

where V^T represents the variance of top- K retrieved texts for a given query image, which is regarded as the uncertainty value of the top-ranking pseudo-pair from text retrieval against this image, and V^I is the variance of top- K retrieved images for a specific query text, represented by the uncertainty value of the top-ranking pseudo-pair from image retrieval against this text.

C. Source-model Adaptation

We use the generated pseudo-pairs as training samples to fine-tune the source model, in which we incorporate uncertainty information into the triplet ranking loss to down-weight the contribution of noisy pairs with high uncertainty, thereby achieving better adaptation results. Concretely, we first map the image and its pseudo-paired text into a common space by two modality-specific networks g and f , respectively. And then the similarity of pseudo-pair is calculated by $S(I, T) = \text{cosine}(f(x_i), g(y_i))$. Finally, the uncertainty-based triplet loss is defined as

$$\mathcal{L}_{triplet}^u = [W_{uncer}^I \cdot (S(I, \hat{T}) - S(I, T)) + m]_+ + [W_{uncer}^T \cdot (S(\hat{I}, T) - S(I, T)) + m]_+, \quad (3)$$

where \hat{T} and \hat{I} denote the negative texts and images in a mini-batch, m is a fixed margin that represents the minimum acceptable distance between the anchor-positive and anchor-negative pairs, and W_{uncer}^I and W_{uncer}^T represent the weights assigned to pseudo-pairs from image to text and from text to image within each mini-batch, respectively. Particularly, each pseudo-pair is assigned a weight based on its corresponding uncertainty, calculated by $W_{uncer}^I = \lambda \exp^{-\beta V^I}$ and $W_{uncer}^T = \lambda \exp^{-\beta V^T}$, where λ and β are adjustable parameters. The allocation of these weights is achieved by making the weight of each pseudo-pair inversely proportional to its corresponding uncertainty. During training, we fine-tune the pre-trained source model using the uncertainty-based triplet loss $\mathcal{L}_{triplet}^u$ for adaptation on generated pseudo-pairs. During testing, the inference process remains as efficient as the source methods, without introducing any additional computations.

III. EXPERIMENTS

A. Implementation details

Datasets. To evaluate the effectiveness of the proposed method, we conduct extensive experiments using three datasets: Flickr30K [23] dataset, COCO [24] dataset and Iaprtc-12 [25] dataset.

Evaluation Metrics. We evaluate the performance of image-text matching by using F1-Score and Recall@ k . The F1-Score balances Precision and Recall, ensuring a comprehensive evaluation without dominance by a single metric. Recall@ k (R@ k) measures the proportion of correct matches within the top- k results, with k set to 1, 5, and 10. The sum of all R@ k values is used to evaluate the overall matching performance, represented as Rsum.

Implementation Details. We construct two source-free adaptation tasks: Iaprtc-12→Flickr30K, Flickr30K→COCO, where Iaprtc-12→Flickr30K represents that the image-text matching model trained on Iaprtc-12 is adapted to Flickr30K, the same to Flickr30K→COCO. For Iaprtc-12→Flickr30K, there exist large domain shifts in both images and texts, due to the difference in both image collection resource and textual description topic between Iaprtc-12 and Flickr30K. For Flickr30K→COCO, there exist significant domain shifts across two modalities, due to the difference in data set size and the number of target categories compared with another adaptation task.

B. Comparisons with state of the arts

To evaluate the adaptation performance of our method, we compare our method with the method that directly applies the source model to the target domain, denoted as VSE++ [26], CAMERA [27], GPO [28], SCAN [29], SGR [30], SAF [30], NUIF [31] and TVRN [32]. To evaluate the effectiveness of the proposed uncertainty learning, we compare our method with the method that fine-tunes the source model using the pseudo-pairs from the target domain without uncertainty-aware learning, denoted as “VSE++ with fine-tuning”, “CAMERA with fine-tuning”, “GPO with fine-tuning”, “SCAN with fine-tuning”, “SGR with fine-tuning”, “SAF with fine-tuning”, “NUIF with fine-tuning” and “TVRN with fine-tuning”.

Table I show the comparison results using the six base models on the Iaprtc-12→Flickr30K and Flickr30K→COCO tasks, respectively. From these tables, we observe that our method outperforms the method that directly applies the source model in all tasks, which demonstrates that our method succeeds in adapting the source model well to target domain by reducing the domain shift.

C. Comparison to pre-trained model

We compare our method with the large pre-trained models CLIP [33], BLIP [34] and OpenCLIP [35], all of which are trained on extensive image-text pairs and exhibit impressive zero-shot generalization capabilities. In our experiments, we utilize the pre-trained CLIP model with a ViT-B/32 backbone, BLIP with a ViT-B/16 backbone, and OpenCLIP with a ViT-H-14378-quickgelu backbone as the source models to generate pseudo-paired training data on the COCO and Flickr30K, which are then used for fine-tuning. From Table II, we observe that our method outperforms others on most evaluation metrics across both datasets, demonstrating the effectiveness of incorporating uncertainty in each pseudo-pair to enhance the adaptability of pre-trained models to new scenarios with different distributions.

D. Ablation Studies

To investigate the effectiveness of each component, we introduce several variants of our method for comparison on Flickr30K→COCO, using the base model VSE++.

Effect of uncertainty in the loss. To evaluate the effectiveness of modeling uncertainty in the triplet loss, we compare our method with the following methods: (1) “w/o W_{uncer}^T ”: removing the uncertainty W_{uncer}^T of the pseudo-pairs that are generated by retrieving texts given an image query; (2) “w/o W_{uncer}^I ”: removing the uncertainty W_{uncer}^I of the pseudo-pairs that are generated by retrieving images given a text query. As reported in Table III, it is obvious that the performances of these two methods degrade, which verifies the benefit of incorporating pseudo-pair uncertainty into both terms in the triplet loss.

Effect of different calculations of uncertainty. To evaluate the effectiveness of using variance to calculate uncertainty in our method, we introduce the following uncertainty calculations for comparison: (1) “uncertainty based on similarity”: the uncertainty of pseudo-pair is calculated by the similarity score between the image and text in the pseudo-pair; (2)

TABLE I: Comparison results of different methods using the six base models on Iaprtc12→Flickr30K and Flickr30K→COCO. The best results are in bold.

Methods	laprtc12→Flickr30K							Flickr30K→COCO								
	Image-to-Text			Text-to-Image			Rsum	F1-Score	Image-to-Text			Text-to-Image			Rsum	F1-Score
	R@1	R@5	R@10	R@1	R@5	R@10			R@1	R@5	R@10	R@1	R@5	R@10		
VSE++ [26]	7.8	20.1	27.1	4.1	13.2	19.3	91.6	35.5	7.4	19.8	29.0	4.7	14.2	21.2	96.4	50.1
VSE++ with fine-tuning	9.6	22.0	30.4	5.0	15.9	22.6	105.5	50.5	7.5	20.9	30.3	5.5	15.6	22.9	102.7	54.4
VSE++ with Ours	10.4	22.3	35.0	5.3	15.7	23.2	108.1	57.0	8.2	21.9	31.3	5.7	16.6	24.3	108.0	62.4
CAMERA [27]	18.3	40.9	52.7	11.4	29.0	38.6	190.9	105.8	26.3	49.5	60.5	18.2	39.2	50.4	244.1	148.3
CAMERA with fine-tuning	23.1	46.6	59.3	15.2	34.8	45.1	224.1	129.9	28.5	53.9	65.1	21.2	44.2	55.5	268.4	166.0
CAMERA with Ours	23.7	46.8	59.6	15.6	35.8	46.2	227.6	134.6	29.5	55.6	67.1	21.6	45.3	56.8	275.8	171.7
GPO [28]	31.4	58.6	68.5	20.7	42.4	52.6	274.2	165.3	30.6	54.5	64.4	20.0	41.3	52.7	263.5	160.9
GPO with fine-tuning	38.7	64.9	74.8	26.8	50.5	61.7	317.4	200.2	35.3	61.7	72.5	25.4	49.7	60.9	305.5	193.1
GPO with Ours	40.9	65.6	76.4	26.5	50.7	62.3	322.5	208.8	36.4	62.6	73.2	25.6	49.6	61.5	309.0	194.8
SCAN [29]	12.5	27.5	39.3	6.2	16.9	23.9	126.3	62.9	15.3	33.7	45.5	14.1	34.1	41.6	181.6	109.2
SCAN with fine-tuning	13.8	31.8	43.6	8.5	22.6	31.9	152.2	79.6	19.0	42.7	55.5	13.9	32.7	43.8	207.5	116.9
SCAN with Ours	14.3	34.1	44.0	9.0	23.7	32.0	157.0	88.4	20.3	43.9	56.7	14.5	33.4	44.3	213.2	121.6
SGR [30]	15.2	32.5	43.8	8.4	19.3	27.1	146.3	75.5	25.4	48.7	59.6	17.7	37.7	48.4	237.6	140.6
SGR with fine-tuning	20.0	41.4	53.7	11.5	27.1	36.4	190.0	102.9	30.4	57.5	69.9	20.2	42.6	54.5	275.1	164.0
SGR with Ours	20.8	42.1	53.5	12.4	28.0	37.1	193.9	106.4	31.8	57.9	70.0	20.4	43.4	54.8	278.3	174.4
SAF [30]	16.4	32.3	43.4	6.8	17.5	24.5	140.9	69.7	21.3	44.2	55.6	15.9	35.1	46.0	218.2	129.0
SAF with fine-tuning	19.8	39.9	50.3	11.4	25.4	34.4	181.2	98.9	29.1	54.9	67.1	18.9	41.1	52.3	263.4	156.0
SAF with Ours	19.8	41.5	52.0	11.3	26.1	35.3	186.0	99.6	30.2	57.1	69.0	19.4	41.7	53.3	270.7	163.9
NUIF [31]	35.9	60.2	69.0	23.5	42.9	52.8	284.3	173.1	28.7	46.0	56.2	23.0	38.5	47.8	240.2	137.4
NUIF with fine-tuning	36.4	64.5	73.4	26.7	44.9	62.5	308.4	194.1	33.7	49.9	61.5	24.6	44.7	51.7	266.1	157.1
NUIF with Ours	38.4	65.1	76.2	26.9	48.9	64.5	320.0	206.5	36.2	54.9	63.8	25.2	45.1	52.6	277.8	174.7
TVRN [32]	24.8	34.8	45.2	14.8	24.0	29.5	173.1	97.3	26.3	45.0	56.1	21.4	37.8	47.7	234.3	135.6
TVRN with fine-tuning	27.8	38.1	46.9	15.7	26.2	29.8	184.5	99.3	29.8	49.4	60.1	23.7	41.6	49.8	254.4	141.3
TVRN with Ours	30.8	43.5	49.0	17.3	28.9	31.4	200.9	110.5	32.3	53.5	63.6	24.8	41.9	51.2	267.0	159.4

TABLE II: Comparison of CLIP, BLIP and OpenCLIP on COCO and Flickr30K dataset. The best results are in bold.

Method	COCO						Flickr30K					
	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	50.2	74.6	83.6	30.4	56.0	66.8	79.0	94.3	98.2	58.0	82.9	89.9
CLIP with fine-tuning	62.6	85.3	91.3	46.6	73.4	82.8	86.7	97.2	98.8	74.3	92.9	96.3
CLIP with Ours	62.2	85.3	92.1	47.5	74.7	83.6	87.6	97.5	98.1	75.2	93.5	96.6
BLIP	57.4	81.1	88.7	41.4	66.0	75.3	76.0	92.8	96.1	58.4	80.0	86.7
BLIP with fine-tuning	64.1	86.8	92.4	49.1	75.7	84.7	84.2	96.1	96.9	74.9	90.5	94.4
BLIP with Ours	65.9	87.1	92.9	51.2	76.0	84.9	83.9	96.7	97.2	75.4	91.4	94.6
OpenCLIP	59.1	82.7	89.8	46.6	72.0	80.6	83.1	95.9	98.7	71.5	91.3	95.0
OpenCLIP with fine-tuning	65.2	87.4	92.8	51.9	75.7	84.0	88.9	98.5	99.2	76.3	94.7	98.0
OpenCLIP with Ours	65.4	87.5	92.6	54.3	76.2	84.7	89.7	98.5	99.3	77.9	94.9	98.5

TABLE III: Results of ablation studies on Flickr30K→COCO task using the base model VSE++. The best results are in bold.

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
w/o W_{uncer}^T	7.7	21.7	30.8	5.6	16.0	23.1
w/o W_{uncer}^I	7.6	21.3	30.3	5.4	16.2	23.5
uncertainty based on similarity	7.1	21.4	30.6	5.6	16.3	24.0
uncertainty without weight	7.7	21.1	31.0	5.7	16.6	23.9
all selection of negative samples	7.3	20.9	30.4	5.6	16.4	24.0
random selection of negative samples	7.3	21.7	30.3	5.6	16.5	24.1
Ours	8.2	21.9	31.3	5.7	16.6	24.3

“uncertainty without weight”: the weight of pseudo-pair is removed to calculate the mean and the variance of pseudo-pair as uncertainty. From Table III, we observe that our method outperforms on most evaluation metrics compared with the two variants, indicating that it is beneficial to mitigate the negative impact of noisy pseudo-pairs by introducing pseudo-pair weights to calculate the uncertainty.

Effect of different strategies of selecting negative samples. To explore which selection strategy is more effective for uncertainty-aware learning, we compare our method with the following variants: (1) “all selection of negative samples”: all the negative samples in a mini-batch are used for training; (2) “random selection of negative samples”: some negative samples are randomly selected from a mini-batch for training. From Table III, it is interesting to notice that our method achieves the best results on most metrics, suggesting that mining harder negative samples with higher uncertainty helps enhance the

positive adaptation.






Original image				
				
A young boy leaning up against a bag of luggage.	A woman is standing on her skis on a slope.	A baby giraffe and a baby zebra stand near a green hut.	People are playing Frisbee or walking near a beach.	A man emphasizing the smartphone he is holding.
Weight of pseudo-pair: 0.3276	Weight of pseudo-pair: 0.4761	Weight of pseudo-pair: 0.2818	Weight of pseudo-pair: 0.0438	Weight of pseudo-pair: 0.0273

Fig. 3: Example of the success and failure cases on Flickr30K→COCO using the base model TVRN.

IV. QUALITATIVE RESULTS

In Fig. 3, we show qualitative results on Flickr30K→COCO using the base model TVRN, the first three are success cases while the last two are failure cases. Successful cases have uncertainty weights about ten times smaller than failure cases, indicating that high-uncertainty pseudo-pairs are assigned smaller weights, reducing their impact during training. By quantifying pseudo-pair uncertainty, our method avoids confirmation bias and self-reinforcing errors.

V. CONCLUSIONS

We have presented a source-free image-text matching approach that adapts a pre-trained model from a source domain to an unlabeled target domain without accessing the source data. To reduce the domain shift from different data distributions, we propose selecting reliable pseudo image-text pairs for model adaptation by calculating variance among retrieved candidates and incorporating uncertainty into the ranking loss to down-weight noisy pairs with high uncertainty. Extensive experiments validate the effectiveness of our method, which is compatible with existing image-text matching approaches.

REFERENCES

- [1] S. Chun, S. J. Oh, R. S. De Rezende, Y. Kalantidis, and D. Larlus, "Probabilistic embeddings for cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8415–8424.
- [2] H. Ma, H. Zhao, Z. Lin, A. Kale, Z. Wang, T. Yu, J. Gu, S. Choudhary, and X. Xie, "Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 051–18 061.
- [3] W. Li, Z. Ma, J. Shi, and X. Fan, "The style transformer with common knowledge optimization for image-text retrieval," *IEEE Signal Processing Letters*, 2023.
- [4] Y. Liu, H. Liu, H. Wang, and M. Liu, "Regularizing visual semantic embedding with contrastive learning for image-text matching," *IEEE Signal Processing Letters*, vol. 29, pp. 1332–1336, 2022.
- [5] H. Lan and P. Zhang, "Learning and integrating multi-level matching features for image-text retrieval," *IEEE Signal Processing Letters*, vol. 29, pp. 374–378, 2022.
- [6] X. Xia, L. Wang, J. Sun, and A. Nakagawa, "Towards generated image provenance analysis via conceptual-similar-guided-slip retrieval," *IEEE Signal Processing Letters*, vol. 31, pp. 1419–1423, 2024.
- [7] W. Zhao, X. Wu, and J. Luo, "Cross-domain image captioning via cross-modal retrieval and model adaptation," *IEEE Transactions on Image Processing*, vol. 30, pp. 1180–1192, 2020.
- [8] C.-W. Kuo and Z. Kira, "Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 969–17 979.
- [9] S. Kornblith, L. Li, Z. Wang, and T. Nguyen, "Guiding image captioning models toward more specific captions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 259–15 269.
- [10] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo, "Promptcap: Prompt-guided image captioning for vqa with gpt-3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2963–2975.
- [11] T. Nguyen, S. Y. Gadre, G. Ilharco, S. Oh, and L. Schmidt, "Improving multimodal datasets with image captioning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [12] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6077–6086.
- [13] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, pp. 3196–3209, 2020.
- [14] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4542–4550.
- [15] M. Parelli, A. Delitzas, N. Hars, G. Vlassis, S. Anagnostidis, G. Bachmann, and T. Hofmann, "Clip-guided vision-language pre-training for question answering in 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5607–5612.
- [16] P. Li, G. Liu, L. Tan, J. Liao, and S. Zhong, "Self-supervised vision-language pretraining for medial visual question answering," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–5.
- [17] Y. Qin, D. Peng, X. Peng, X. Wang, and P. Hu, "Deep evidential learning with noisy correspondence for cross-modal retrieval," in *Proceedings of the ACM International Conference on Multimedia*, 2022, pp. 4948–4956.
- [18] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, and X. Peng, "Learning with noisy correspondence for cross-modal matching," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 406–29 419, 2021.
- [19] J. V. Zidek and C. Van Eeden, "Uncertainty, entropy, variance and the effect of partial information," *Lecture Notes-Monograph Series*, vol. 42, pp. 155–167, 2003.
- [20] J. Chen, C. van Eeden, and J. Zidek, "Uncertainty and the conditional variance," *Statistics & probability letters*, vol. 80, pp. 1764–1770, 2010.
- [21] A. Loquercio, M. Segu, and D. Scaramuzza, "A general framework for uncertainty estimation in deep learning," *IEEE Robotics and Automation Letters*, vol. 5, pp. 3153–3160, 2020.
- [22] X. Xia, T. Liu, B. Han, M. Gong, J. Yu, G. Niu, and M. Sugiyama, "Sample selection with uncertainty of losses for learning with noisy labels," in *International Conference on Learning Representations (ICLR)*, 2022, pp. 1–23.
- [23] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.
- [24] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.
- [25] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," in *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2020, pp. 647–664.
- [26] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: improving visual-semantic embeddings with hard negatives," in *British Machine Vision Conference (BMVC)*, 2018, p. 12.
- [27] L. Qu, M. Liu, D. Cao, L. Nie, and Q. Tian, "Context-aware multi-view summarization network for image-text matching," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 1047–1055.
- [28] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the best pooling strategy for visual semantic embedding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021.
- [29] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 201–216.
- [30] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.
- [31] H. Zhang, L. Zhang, K. Zhang, and Z. Mao, "Identification of necessary semantic undertakers in the causal view for image-text matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7105–7114.
- [32] S. Pang, Y. Zeng, J. Zhao, and J. Xue, "A mutually textual and visual refinement network for image-text matching," *IEEE Transactions on Multimedia*, 2024.
- [33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [34] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.
- [35] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," July 2021, if you use this software, please cite it as below. [Online]. Available: <https://doi.org/10.5281/zenodo.5143773>