



ItrievalKD: An Iterative Retrieval Framework Assisted with Knowledge Distillation for Noisy Text-to-Image Retrieval

Zhen Liu, Yongxin Zhu, Zhujin Gao, Xin Sheng, and Linli Xu^(✉)

State Key Laboratory of Cognitive Intelligence,
University of Science and Technology of China, Hefei, China
{liuzhenz, zyx2016, gaozhujin, xins}@mail.ustc.edu.cn, linlixu@ustc.edu.cn

Abstract. Benefiting from the superiority of the pretraining paradigm on large-scale multi-modal data, current cross-modal pretrained models (such as CLIP) have shown excellent performance on text-to-image retrieval. However, the current research mainly focuses on the scenarios with strong matching of images and texts, which is not always available in practice. For example, in social media content or daily communication, the text is not always completely related to the image and may also contain some irrelevant content, which introduces non-negligible noise to text-to-image retrieval. The noisy multi-modal setting is significantly different from the current cross-modal pretraining corpus, which may lead to significant degradation of the retrieval performance of the general image-text retrieval models. In this paper, we focus on the task of noisy text-to-image retrieval and propose an iterative retrieval framework which firstly retrieves the key-semantic information from the noisy text with knowledge distillation, followed by retrieving the relevant image from the image pool with the key-semantic clue. Experiments on Noisy-MS-COCO and PhotoChat datasets confirm the superiority of the proposed iterative retrieval framework in the task of noisy text-to-image retrieval compared with the general retrieval models.

Keywords: Image-text retrieval · Knowledge distillation · Extractive summarization

1 Introduction

The task of cross-modal image-text retrieval is to retrieve samples from one modal with the guidance of the samples from the other modal. With the rapid growth of the data from various modalities, cross-modal image-text retrieval has a wide range of applications, helping users quickly locate data from a specific modal that is relevant to the current query. In general, cross-modal image-text retrieval consists of two sub-tasks, which are image-to-text (I2T) and text-to-image (T2I) retrieval respectively. In this paper, we focus on T2I retrieval which aims at retrieving the most relevant image according to the textual context.

As a cross-modal task, the major challenge of T2I retrieval is how to bridge the semantic gap between different modals. To tackle that, the transformer-based cross-modal pretraining models have been successfully applied to various tasks. By pretraining on large-scale image-text datasets, different modalities are encoded into a common semantic space, yielding modal-agnostic semantic representations. CLIP [1] is one of the most representative models, the zero-shot retrieval performance of which on the commonly-used image-text dataset MSCOCO [12] is competitive with the finetuning performance of the previous pretraining models.

Despite the outstanding performance of the methods based on large-scale cross-modal pretraining, it is often overlooked that most of the current research on T2I retrieval assumes that the query-key data is strongly correlated. Nevertheless, the image-text matching relationship may be weakly correlated in real scenarios. For example, on social media platforms such as Twitter, people tend to share their daily life in a combination of images and texts, where the text may involve some content that is irrelevant to the image. As a matter of fact, the image-text scenarios can be very noisy in practice. Figure 1 shows an example in the PhotoChat dataset from the photo sharing task [19]. The motivation of the photo sharing task is the popularity of photo-sharing in online chat, the goal of which is to retrieve the corresponding image of the conversational context, most of which is irrelevant chat. In this case, applying the CLIP model directly fails to retrieve the correct image, as the CLIP model obviously ignores the information of “strawberry” and “blueberry” in the dialogue.

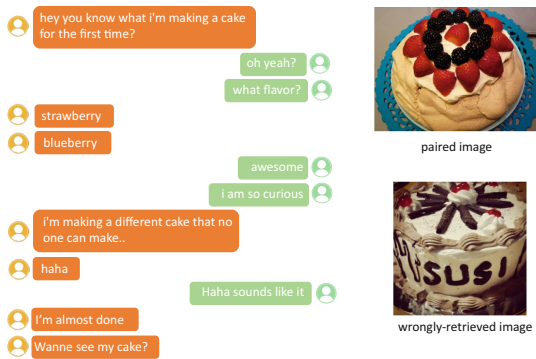


Fig. 1. An example in the PhotoChat dataset. The top right image is the image relevant to the left dialog context, and the bottom right image is retrieved with the CLIP model.

To further analyze why the CLIP model fails to retrieve the correct image in the above example, we conduct a simple investigation. Specifically, we construct a simple noisy T2I scenario named Noisy-MSCOCO by injecting noise to the dataset MSCOCO. We generate some noisy sentences according to the captions from MSCOCO and then mix them to get the final noisy text. Figure 2 shows the

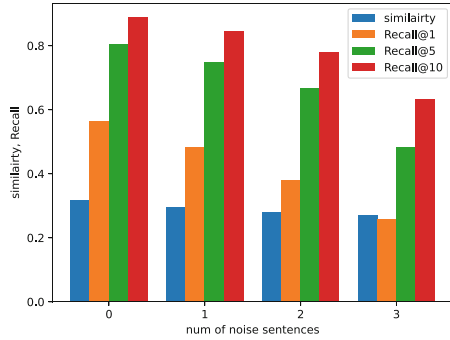


Fig. 2. The retrieval results and the similarity between image-text pairs on Noisy-MS-COCO given the number of noisy sentences. The results are reported in the zero-shot setting with CLIP.

retrieval performance based on CLIP as the number of noisy sentences grows. We also report the cosine similarity, i.e. the image-text alignment score between the representations of the two modalities calculated by CLIP. It can be seen that with the increasing noise, the image-text similarity between the text and the corresponding image decreases as expected, which results in a significant degradation in the retrieval performance. This poses an interesting problem in the retrieval task, which is motivated to capture the relevant information from raw data, whereas the noisy sentences are obviously harmful to the T2I retrieval performance.

In this paper, we focus on text-to-image retrieval where the text may contain a lot of noise, and define the problem as noisy text-to-image retrieval (NT2I). We propose an iterative retrieval framework assisted with knowledge distillation (ItrievalKD). Essentially, to alleviate the influence of irrelevant textual content on the retrieval performance, it is necessary to extract the image-related content from the noisy text as the key-semantic text. Unfortunately, the supervision information of key-semantic text is not available for training the extractor in most cases. Therefore in our iterative retrieval framework, we start with exploiting CLIP to obtain the key-semantic annotations, and then proceed to retrieve the relevant image from the image pool with the key-semantic clue. Furthermore, due to the lack of image annotations during testing, we propose to adopt knowledge distillation to distill the image-text matching knowledge from the cross-modal model CLIP to the plain-text model BERT [3], which can be used to obtain the key-semantic information in the testing phase. The relevant image can then be retrieved with the key-semantic clue from the image pool.

In summary, the main contributions of the work are:

- We propose an iterative retrieval framework for the noisy T2I task, where the text contains noise in text-to-image retrieval.

- We adopt knowledge distillation to transfer the image-text matching knowledge from the cross-modal model to the plain-text model to alleviate the lack of image annotations in the testing stage.
- The experimental results on the Noisy-MSCOCO and PhotoChat datasets demonstrate the superiority of the proposed method.

2 Related Work

Most early cross-modal retrieval methods adopt separate encoders to encode images and texts respectively [18]. While being efficient, these independent feature-encoding models usually produce sub-optimal performance due to the lack of interactions between modals. [11] is the first attempt to consider the dense pairwise cross-modal interactions which achieves tremendous accuracy improvements. After that, various cross-modal interaction methods [2, 8] have been proposed to extract the features of both text and image. On the other hand, methods with only global cross-modal are restricted in the sense that text descriptions usually contain fine-grained correlations with images, which are easily smoothed by global alignment. To address that, some works [7, 17] propose to explore the region (or patch) to word correspondences. An alternative solution is the pretrain-then-finetune paradigm driven by the global alignment method [1, 9], which can achieve satisfactory results with improved robustness, with the help of the large-scale pretraining data.

3 Methodology

In this section, we elaborate on the iterative retrieval framework for the noisy text-to-image retrieval task. We start with the problem definition and a brief overview of the CLIP model which is employed in the iterative retrieval process. Then the architecture of the proposed model will be described in detail.

Problem Definition. Given a parallel image-text dataset (T, V) , each sample pair consists of a noisy text t_i and a relevant image v_i , where $t_i = \{t_i^1, t_i^2, \dots, t_i^k, \dots, t_i^m\}$ is composed of multiple sentences and t_i^k represents the k -th sentence. The task is to retrieve the most relevant image v_i to the noisy text t_i from the image pool V with the proposed iterative retrieval model $R(t_i, V)$.

3.1 Preliminaries: CLIP

CLIP is trained to learn visual representations with natural language supervision. As shown in Fig. 3, it consists of a text encoder \mathbb{T} which is a GPT [15] style Transformer model, and an image encoder \mathbb{V} which can be either a Vision Transformer (ViT) [4] or a Residual Convolutional Neural Network (ResNet) [5]. Then the dot product between the two outputs of the above two encoders will be used as the alignment score of the input image and the text. The model is

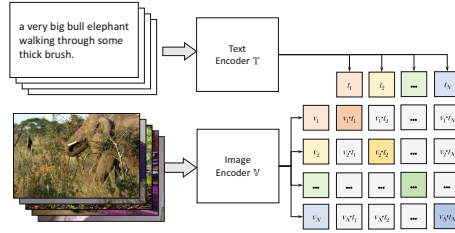


Fig. 3. The framework of the CLIP model.

pretrained to distinguish aligned image-text pairs from randomly combined ones with a contrastive loss,

$$\mathcal{L}_{\text{NCE}} = - \left(\log \frac{\exp(\text{sim}(v_i, t_i)/\alpha)}{\sum_j \exp(\text{sim}(v_i, t_j)/\alpha)} + \log \frac{\exp(\text{sim}(t_i, v_i)/\alpha)}{\sum_j \exp(\text{sim}(t_i, v_j)/\alpha)} \right) \tag{1}$$

where α is the temperature coefficient to be learned in CLIP. The image-text alignment score $\text{sim}(v_i, t_i)$, which is the similarity mentioned above, is calculated as follows,

$$\text{sim}(v_i, t_i) = \frac{\mathbb{T}(t_i) * \mathbb{V}(v_i)}{\|\mathbb{T}(t_i)\|^2 * \|\mathbb{V}(v_i)\|^2} \tag{2}$$

3.2 Model Architecture

The overall framework of the proposed model is shown in Fig. 4, which is compared to the general method in the left panel. Instead of directly taking the noisy text as the query to retrieve the most relevant image from the image pool, which may degrade the retrieval performance as discussed above, our proposed method ItrievalKD first extracts the key-semantic information from the noisy text to alleviate the influence of irrelevant textual content on the retrieval performance, followed by retrieving the relevant image according to the key-semantic clue. Below we will describe the iterative retrieval framework in detail.

Retrieving the Key-Semantic Text in the Noisy Text. Due to the lack of ground truth regarding the key-semantic annotations in most NT2I cases as supervision, it is necessary to retrieve the key-semantic content in the noisy text at first. Here we propose a simple yet effective annotation strategy. We consider each sentence as a basic unit of semantic information in the noisy text. In the NT2I scenario, the key-semantic content in the noisy text should be highly-related to the corresponding image, and the rest should be irrelevant. Hence, the choice of key-semantic text heavily depends on the corresponding

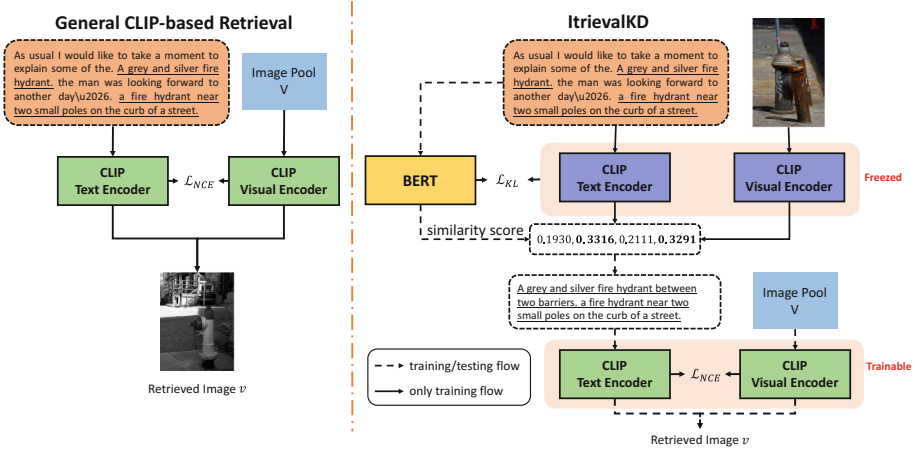


Fig. 4. An illustration of the general retrieval method with the CLIP model (the left panel) and the proposed ItrievalKD framework (the right panel). The underlined part of the input text corresponds to the key-semantic content, and the rest are noisy sentences.

image. Therefore, we calculate the score s_i^k for each sentence t_i^k using Eq. (2) as $s_i^k = sim(v_i, t_i^k)$, which represents the similarity score between the image v_i and the sentence t_i^k . The higher the score is, the more likely the sentence is key-semantic to the image. We then take κ sentences with the highest scores as the key-semantic sentences \hat{t}_i of the noisy text t_i .

Knowledge Distillation for Key-Semantic Extraction. Nevertheless, the lack of the image information paired with the noisy text makes it impossible to directly apply the above strategy to select the key-semantic content during the testing stage, when only the noisy text is available. To resolve that, we need to transfer the image-text correlation knowledge of the CLIP model to a plain-text model, based on which the key-semantic text can be obtained from the plain-text model during the testing stage.

Specifically, we adopt the Knowledge Distillation (KD) technique [6] to distill the knowledge of the image-text content relevance from the teacher model (i.e., CLIP) to the student model (i.e., BERT). The student model BERT is required to mimic the behaviors of the teacher network CLIP when calculating the image-text content relevance scores, followed by ranking the sentences according to the scores.

Since the BERT model picks sentences in unit of sentence, we follow the same input form as BERTSUM [13], which is a method for the extractive summarization task. We insert a [CLS] token before each sentence and a [SEP] token after each sentence. Interval segment embedding is used to distinguish multiple sentences within a text. Finally, we obtain the input to the BERT model by combining the token embeddings, interval segment embeddings and position embeddings. The vector h_i^k , which is the corresponding vector of the k -th [CLS] token

from the top BERT layer, will be used as the representation of the sentence t_i^k .

After obtaining the sentence representation h_i^k from BERT, we build a linear layer and a sigmoid layer on the top of the BERT outputs to learn the sentence-image matching scores,

$$\hat{s}_i^k = \sigma(Wh_i^k + b) \quad (3)$$

where σ is the activation function (sigmoid in this work).

We use the Kullback-Leibler (KL) divergence [10] to quantify the discrepancy between the ranking score distributions of the plain-text model BERT and the multi-modal model CLIP. Via knowledge distillation, the plain-text model BERT directly imitates the score distribution from the teacher model CLIP. Formally, the training objective is to minimize the following loss functions with temperature τ ,

$$\begin{aligned} \mathcal{L}_{\text{KL}} &= -p \ln \frac{q}{p} \\ p(\hat{s}_i^k, \tau) &= \frac{\exp(\hat{s}_i^k/\tau)}{\sum_k \exp(\hat{s}_i^k/\tau)} \\ q(s_i^k, \tau) &= \frac{\exp(s_i^k/\tau)}{\sum_k \exp(s_i^k/\tau)} \end{aligned} \quad (4)$$

Image Retrieval with the Key-Semantic Clue. After obtaining the key-semantic text \hat{t}_i according to the scores, we can take it instead of the noisy text t_i for cross-modal retrieval. We can adopt Eq. (1) to finetune the CLIP model to further augment the performance.

3.3 Training and Inference

Training. The BERT model is trained with the knowledge distilled from the CLIP model by minimizing \mathcal{L}_{KL} , while the parameters of the CLIP model are frozen. It is optional to finetune the CLIP model with \mathcal{L}_{NCE} for further performance when retrieving the relevant image with the key-semantic clue. We report the performance both in the zero-shot setting and finetuning setting. As the CLIP model is prone to overfitting when finetuning, we use the noisy text for training to alleviate this problem in the finetuning setting.

Inference. We first retrieve the key-semantic text from the noisy text with the BERT model, and proceed to retrieve the relevant image from the image pool with the key-semantic clue.

4 Experiments

4.1 Datasets

Noisy-MSCOCO. Given the lack of available datasets in the NT2I scenario, we extend the MSCOCO dataset with additional noise to construct the

Noisy-MSCO dataset. Specifically, we randomly select 10,000, 1000 and 1000 image-text pairs from MSCOCO for training, validation and test respectively. Noisy sentences are generated with the GPT-2 [16] model by extending each caption with a prompt “and” where the maximum length is set as 35. Finally, we randomly sample n_{key} and n_{noise} sentences from the original captions and noisy sentences separately, followed by shuffling them to construct the Noisy-MSCO dataset. In practice, n_{key} is set to 3, and n_{noise} is selected from $\{0, 1, 2, 3\}$.

PhotoChat. PhotoChat is a multi-modal conversation dataset, where each dialogue is paired with an image that is shared during the conversation. Following previous works, we only consider the conversation content of the party who sends the image, because only this party can see the image before sending it.

4.2 Evaluation Metrics

We use Recall@K (R@K), computed as “the fraction of times a correct item was found among the top K results” as the evaluation metric. Specifically, we choose R@1, R@5, and R@10, as well as the sum of them which we denote as “SUM” as [19] to evaluate the proposed method.

4.3 Implementation Details

The proposed model mainly consists of modules based on BERT and CLIP. For BERT, we adopt the “bert-base-uncased” version. We set the batch size to 32, the maximum input length to 256 and the temperature coefficient τ in Equation (4) to 1. During the validation and test stages, for the Noisy-MSCO dataset, we directly adopt n_{key} as κ which is the number of key sentences extracted; and for the PhotoChat dataset, we set κ to 3 in the zero-shot setting and 4 in the finetuning setting as this work best. The best BERT model is chosen according to the accuracy in predicting the key-semantic sentences on the validation set. We employ CLIP (ViT-B/32) and CLIP (RN50) from the series of the CLIP models, and set CLIP (ViT-B/32) as the default. During finetuning, the batch sizes of CLIP (ViT-B/32) and CLIP (RN50) are set as 128 and 64 respectively, and we scale the max input length of the CLIP model to 128 as the original CLIP model limits the text input length to 77 which may be exceeded by the text length in the PhotoChat dataset. The random seed is set to 1 and the Adam optimizer is employed with the learning rate of $1e - 5$.

4.4 Baselines

We mainly compare the proposed framework with the general CLIP-based retrieval model. In addition, since the stage of extracting key sentences from the noisy text is similar to the extractive summarization task, we also select two classical unsupervised extractive summarization methods: 1) TF-IDF [10], a statistical method used to assess the importance of words in a document of a

Table 1. The zero-shot retrieval results on the Noisy-MSCOCO dataset. Key-CLIP, CLIP and ItrievalKD correspond to CLIP retrieval with the ground truth key-sentences, with the noisy text and iterative retrieval with the predicted key-sentences respectively.

n_{noise}	CLIP (ViT-B/32)	zero-shot for CLIP				CLIP (RN50)	zero-shot for CLIP			
		R@1	R@5	R@10	SUM		R@1	R@5	R@10	SUM
0	Key-CLIP	56.3	80.5	88.8	225.6	Key-CLIP	55.2	79.9	87.3	222.4
1	CLIP	48.4	74.7	84.5	207.6	CLIP	49.7	73.8	83.1	206.6
	ItrievalKD	52.9	80.4	88.8	222.1	ItrievalKD	53.7	79.0	87.4	220.1
2	CLIP	37.9	66.7	78.0	182.6	CLIP	44.6	70.8	80.3	195.7
	ItrievalKD	53.6	80.4	88.4	222.4	ItrievalKD	54.0	79.5	88.1	221.6
3	CLIP	25.7	48.4	63.4	137.5	CLIP	32.3	59.1	71.3	162.7
	ItrievalKD	53.9	79.8	88.4	222.1	ItrievalKD	54.0	77.7	86.4	218.1

Table 2. The zero-shot and finetuning retrieval results on the PhotoChat dataset.

CLIP version	model	zero-shot for CLIP				finetuning for CLIP			
		R@1	R@5	R@10	SUM	R@1	R@5	R@10	SUM
CLIP (ViT-B/32)	CLIP	23.3	42.9	52.3	118.5	38.5	64.0	72.3	174.8
	TF-IDF-CLIP	13.1	27.2	35.2	75.5	27.2	49.3	57.8	134.3
	TextRank-CLIP	12.8	27.9	35.7	76.4	22.5	42.5	50.9	115.9
	ItrievalKD	26.7	46.3	55.5	127.6	41.2	64.0	72.1	177.3
CLIP (RN50)	CLIP	25.8	43.6	52.0	121.4	31.6	58.7	67.6	157.9
	TF-IDF-CLIP	13.5	27.1	34.4	75	24.1	43.7	54.2	122.0
	TextRank-CLIP	10.6	20.6	27.2	58.4	19.0	37.1	47.9	104.0
	ItrievalKD	26.3	45.2	55.6	127.1	34.5	59.3	68.8	162.6

corpus. Specifically, we take the maximum TF-IDF value of the words in a sentence as the importance score of the sentence; 2) TextRank [14], a graph-based ranking algorithm, in which we construct the graph by treating each sentence as a node. We extract the key sentences from the noisy text with the above two extractive summarization methods, based on which we retrieve the relevant image, which are named as TF-IDF-CLIP and TextRank-CLIP.

4.5 Retrieval Results

The retrieval results on the Noisy-MSCOCO dataset are shown in Table 1. As CLIP is prone to overfitting on the MSCOCO dataset, we only report the results in the zero-shot setting. In the experiments, we compare the retrieval performance of the CLIP model with the ground truth annotations (Key-CLIP) to the one with the noisy text as the query, where we can observe that the retrieval performance of CLIP degrades significantly with the noise increases, compared to the model with no noise. In comparison, the proposed method ItrievalKD can effectively eliminate the influence of the noisy sentences by extracting the

Table 3. The accuracy in retrieving key sentences in the noisy text with the sentence-image matching scores calculated by CLIP on the Noisy-MSCOCO dataset.

n_{noise}	CLIP (ViT-B/32)	CLIP (RN50)
1	0.9823	0.9843
2	0.9760	0.9793
3	0.9653	0.9643

key-semantic content from the noisy text and achieve comparable results with the noise-free performance of Key-CLIP. For example, R@1 drops from 56.3 to 25.7 when n_{noise} increases to 3 in the zero-shot setting with CLIP (ViT-B/32), while reaching 53.9 when ItrievalKD is applied.

Table 2 shows the zero-shot and finetuning results on the PhotoChat dataset. The retrieval results of the proposed ItrievalKD surpasses the CLIP model in both zero-shot and finetuning settings, demonstrating its effectiveness. Especially, SUM increases from 118.5 to 127.6 in the zero-shot setting over CLIP (ViT-B/32). In addition, it can be observed that both of the two unsupervised summarization methods (i.e., TF-IDF-CLIP and TextRank-CLIP) even degrade the retrieval performance, which implies that the conventional unsupervised summarization methods are not suitable for key-semantic extraction in the NT2I task.

The results of the ItrievalKD framework based on CLIP (ViT-B/32) and CLIP (RN50) follow the similar trend, which demonstrates the effectiveness and robustness of the proposed method.



4.6 The Effectiveness of Retrieving the Key-Semantic Text in the Noisy Text with CLIP

We proceed to verify the effectiveness of retrieving the key-semantic sentences in the noisy text with CLIP. We show the performance of adopting CLIP to retrieve key sentences on the Noisy-MSCOCO dataset in Table 3. As the Noisy-MSCOCO dataset has key-sentence labels, we use accuracy to evaluate the performance of retrieving key sentences by CLIP. It is observed that, although the accuracy decreases slightly with the noise increases, the accuracy over CLIP (ViT-B/32) on the Noisy-MSCOCO dataset remains 96.53% even when n_{noise} is set to 3. It validates that the strong ability of retrieving the key sentences from the noisy text enables ItrievalKD to achieve comparable results with the noise-free performance of Key-CLIP as shown in Table 1.

4.7 Case Study

An example on the Noisy-MSCOCO dataset is given in Table 4. The general retrieval method given the entire noisy text as the query would return the wrong image, while the ItrievalKD method can retrieve the truly-relevant image. In this

Table 4. Case study on the Noisy-MS-COCO dataset in the zero-shot setting over CLIP. The underlined department of the text is the key-semantic clue.

Text	Negative Image
<p>at 9:46 p.m. this morning, someone in Florida called the police to report that I am extremely excited to present the 6th edition of The Game, a collection of the a red fire hydrant near a dirt road with trees in the <u>background</u></p> <p><u>A red fire hydrant in a forest setting.</u></p> <p><u>A close of a red fire hydrant next to a road</u></p> <p>In his final days in office, President Barack Obama has put his administration on a high alert. The</p>	
	<p data-bbox="818 416 992 441" style="text-align: center;">Positive Image</p> 

case, if the noisy text is used, the general retrieval model may pay attention to the key information “fire hydrant” while ignoring the details such as “forest”, “tree”, and “in the background”. By capturing the key-semantic information in the noisy text, the proposed method can avoid this problem.

5 Conclusion

In this paper, we propose an iterative retrieval framework assisted with knowledge distillation ItrievalKD for the text-to-image retrieval task when the query text contains noise unrelated to the relevant image. As the irrelevant information in the text is harmful to the capturing of key-semantic part for the general retrieval model, the proposed method ItrievalKD first obtains the key-semantic information from the noisy text, followed by retrieving the relevant image from the image pool based on the key-semantic clue. We verify the effectiveness of the proposed method on the Noisy-MS-COCO and PhotoChat datasets.

Acknowledgments. This research was supported by the National Key Research and Development Program of China (Grant No. 2022YFB3103100), the National Natural Science Foundation of China (Grant No. 62276245), and Anhui Provincial Natural Science Foundation (Grant No. 2008085J31).

References

1. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
2. Cui, Y., et al.: Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 797–806 (2021)

3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *Comput. Sci.* **14**(7), 38–39 (2015)
7. Ji, Z., Chen, K., Wang, H.: Step-wise hierarchical alignment network for image-text matching. arXiv preprint [arXiv:2106.06509](https://arxiv.org/abs/2106.06509) (2021)
8. Ji, Z., Wang, H., Han, J., Pang, Y.: Saliency-guided attention network for image-sentence matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5754–5763 (2019)
9. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, pp. 4904–4916. PMLR (2021)
10. Kullback, S., Leibler, R.A.: On information and sufficiency. *Ann. Math. Stat.* **22**(1), 79–86 (1951)
11. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European conference computer vision (ECCV), pp. 201–216 (2018)
12. Lin, T.-Y., et al.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
13. Liu, Y.: Fine-tune bert for extractive summarization (2019)
14. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text (2004)
15. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. *OpenAI blog* **1**(8), 9 (2019)
17. Wu, H., et al.: Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6609–6618 (2019)
18. Wu, Y., Wang, S., Song, G., Huang, Q.: Learning fragment self-attention embeddings for image-text matching. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2088–2096 (2019)
19. Zang, X., Liu, L., Wang, M., Song, Y., Zhang, H., Chen, J.: Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. arXiv preprint [arXiv:2108.01453](https://arxiv.org/abs/2108.01453) (2021)