Image-text Retrieval: A Survey on Recent Research and Development

Min Cao[1] , **Shiping Li**[1] , **Juntao Li**[1] , **Liqiang Nie**[2] and **Min Zhang**[1,3]
[1]Soochow University
[2] Shandong University
[3] Harbin Institute of Technology, Shenzhen
{mcao,ljt,minzhang}@suda.edu.cn, spli@stu.suda.edu.cn, nieliqiang@gmail.com

Abstract

In the past few years, cross-modal image-text retrieval (ITR) has experienced increased interest in the research community due to its excellent research value and broad real-world application. It is designed for the scenarios where the queries are from one modality and the retrieval galleries from another modality. This paper presents a comprehensive and up-to-date survey on the ITR approaches from four perspectives. By dissecting an ITR system into two processes: feature extraction and feature alignment, we summarize the recent advance of the ITR approaches from these two perspectives. On top of this, the efficiency-focused study on the ITR system is introduced as the third perspective. To keep pace with the times, we also provide a pioneering overview of the cross-modal pre-training ITR approaches as the fourth perspective. Finally, we outline the common benchmark datasets and evaluation metric for ITR, and conduct the accuracy comparison among the representative ITR approaches. Some critical yet less studied issues are discussed at the end of the paper.

1 Introduction

Cross-modal image-text retrieval (ITR) is to retrieve the relevant samples from one modality as per the given user expressed in another modality, usually including two sub-tasks: image-to-text (i2t) and text-to-image (t2i) retrieval. ITR has extensive application prospects in the search field and is a valuable research topic. Thanks to the prosperity of deep models for language and vision, we have witnessed the great success of ITR in the past few years [Frome *et al.*, 2013; Li *et al.*, 2021]. For instance, along with the rising of BERT, the transformer-based cross-modal pre-training paradigm has gained momentum, and its pretrain-then-finetune form has been extended to the downstream ITR task, accelerating its development.

It is worth mentioning that several prior efforts have been dedicated to conduct a survey on ITR. They, however, suffer from the following two limitations: 1) Beyond the ITR task, other multi-modal tasks, such as video-text retrieval and visual question answering, are also explored, resulting in a less in-depth ITR survey [Uppal *et al.*, 2022; Baltrušaitis *et al.*, 2018]; 2) the pre-training paradigm is largely untapped in existing surveys [Chen *et al.*, 2020b] and is indeed the mainstream nowadays. In light of these, we conduct a comprehensive and up-to-date survey on the ITR task in this paper, especially with an investigation on the pre-training paradigm.

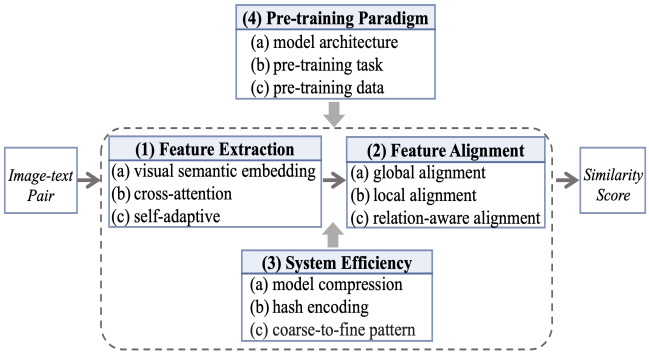

Figure 1: Illustration of the classification skeleton of ITR approaches from four perspectives.

An ITR system generally consists of the feature extraction process with the image/text processing branches and the feature alignment process with an integration module. Contextualized in such an ITR system, we construct taxonomy from four perspectives to overview ITR approaches. Figure 1 illustrates the classification skeleton for the ITR approaches.

(1) Feature extraction. Existing approaches on extracting discriminative image and text features fall into three categories. 1) The visual semantic embedding based approaches work towards learning features independently. 2) The cross-attention approaches, by contrast, learn features interactively. And 3) the self-adaptive approaches aim at learning features with self-adaptive patterns.

(2) Feature alignment. The heterogeneity of multimodal data makes the integration module important for aligning image and text features. Existing approaches are in two variants. 1) The global alignment-driven approaches align global features across modalities. 2) Beyond that, some approaches attempt to find local alignment explicitly at a fine-grained level, so-called the local alignment-involved approaches.

(3) System efficiency. Efficiency plays a pivotal role in an excellent ITR system. Apart from the research on improving ITR accuracy, a stream of works pursues a high-efficient retrieval system in three different ways. 1) The hash encoding approaches reduce the computational cost via binarizing the features in float format. 2) The model compression approaches emphasize low-energy and lightweight running. And 3) the fast-then-slow approaches perform the retrieval via a coarse-grained fast retrieval first and then a fine-grained slow one.

(4) Pre-training paradigm. To stand at the forefront of research development, we also gain insights into the cross-modal pre-training approaches for the ITR task, which has gained much attention recently. Compared with the conventional ITR[1], the pre-training ITR approaches benefit from the rich knowledge that is implicitly encoded by the large-scale cross-modal pre-trained models, yielding encouraging performance even without sophisticated retrieval engineering. In the context of the ITR task, the cross-modal pre-training approaches are still applied to the taxonomy of the above three perspectives. However, to characterize the pre-training ITR approaches more clearly, we re-classify them by three dimensions: model architecture, pre-training task, and pre-training data.

In what follows, we summarize the ITR approaches based on the above taxonomy of the first three perspectives in Section 2, and make a particular reference to the pre-training ITR approaches, i.e., the fourth perspective, in Section 3. A detailed overview of the common datasets, evaluation metric and accuracy comparison among representative approaches is presented in Section 4, followed by the conclusion and future work in Section 5.

2 Image-text Retrieval

2.1 Feature Extraction

Extracting the image and text features is the first and also the most crucial process in the ITR system. Under three different developing trends: visual semantic embedding, cross-attention and self-adaptive, as shown in Figure 2, the feature extraction technology in ITR is thriving.

Visual semantic embedding (VSE). Encoding image and text features independently is an intuitive and straightforward way for ITR, which was firstly proposed in [Frome *et al.*, 2013]. Afterwards, such VSE based approaches are widely developed in roughly two aspects. 1) In terms of data, a stream of works [Wu *et al.*, 2019b; Chun *et al.*, 2021] tries to excavate the high-order data information for learning powerful features. They learn features with equal treatment for all data pairs. In contrast, some researchers [Faghri *et al.*, 2017] propose to weight the informative pairs to improve the discrimination of features, and others [Hu *et al.*, 2021] pay more attention to the mismatched noise correspondences in data pairs for the feature extraction. Recently, riding the wave

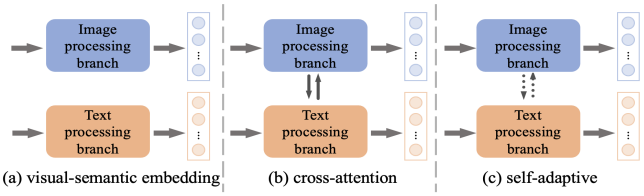

Figure 2: Illustration of different feature extraction architectures.

of the large-scale cross-modal pre-training technology, some works [Jia *et al.*, 2021; Huo *et al.*, 2021] leverage large-scale web data directly to pre-train the image and text feature extractors, exhibiting the impressive performance on the downstream ITR task. 2) Regarding the loss function, the ranking loss is commonly used in the VSE based approaches [Frome *et al.*, 2013; Faghri *et al.*, 2017] and constrains the inter-modal data relationship for learning features. Besides that, [Wang *et al.*, 2018] proposed a maximum-margin ranking loss with the neighborhood constraints for better extracting features. [Zheng *et al.*, 2020] proposed an instance loss explicitly considering the intra-modal data distribution.

Owing to the independent feature encoding, the VSE based approach enables a high-efficiency ITR system in which the features of massive gallery samples can be pre-computed offline. However, it may bring suboptimal features and limited ITR performance due to less exploration of the interaction between the image and text data.

Cross-attention (CA). [Lee *et al.*, 2018] made the first attempt to consider the dense pairwise cross-modal interaction and yielded tremendous accuracy improvements at the time. Since then, various CA approaches have been put forward to extract features. Employing the transformer architecture, the researchers can simply operate on a concatenation of image and text to the transformer architecture, thereby learning the cross-modal contextualized features. It opens up a rich line of studies on the transformer-like CA approach [Lu *et al.*, 2019; Chen *et al.*, 2020c]. Moreover, injecting some additional contents or operations into the cross-attention for assisting the feature extraction is also a new line of research. [Ji *et al.*, 2019] adopted a visual saliency detection module to guide the cross-modal correlation. [Cui *et al.*, 2021] integrated intra- and cross-modal knowledge to learn the image and text features jointly.

The CA approach narrows the data heterogeneity gap and tends to obtain high-accuracy retrieval results, yet comes at a prohibitive cost since each image-text pair must be fed into the cross-attention module online.

Self-adaptive (SA). Instead of a fixed computation flow for extracting features in the VSE based and CA approaches, [Qu *et al.*, 2021] started from scratch and educed a self-adaptive modality interaction network in which different pairs can be adaptively inputted into different feature extraction mechanisms. It powerfully inherits the respective merits of the above two groups and is classified as the SA approach.

2.2 Feature Alignment

After the feature extraction, it is desirable to align cross-modal features to compute pairwise similarity and achieve

[1]We denote the ITR approaches without the benefit from cross-modal pre-training knowledge as the conventional ITR for distinguishing them from the ITR approaches under the cross-modal pre-training paradigm.

retrieval. The global and local alignments are two directions.

Global alignment. In the global alignment-driven approach, the image and text are matched from a global viewpoint, as shown in Figure 3 (a). Early works [Faghri *et al.*, 2017; Wang *et al.*, 2018] are usually equipped with a clear and simple two-stream global feature learning network, and the pairwise similarity is computed by the comparison between global features. Later studies [Sarafianos *et al.*, 2019; Zheng *et al.*, 2020] focus on improving such two-stream network architecture for better aligning global features. Nonetheless, the above approaches with only the global alignment always present limited performance since the textual description usually contains finer-grained detail of image, which is prone to be smoothed by the global alignment. However, there is an exception. The recent global alignment-driven approaches in a pretrain-then-finetune paradigm [Jia *et al.*, 2021] tend to produce satisfactory results, attributed to the enlarging scale of pre-training data.

All in all, only applying the global alignment to ITR could lead to a deficiency of the fine-grained correspondence modeling and is relatively weak for computing reliable pairwise similarity. Considering the alignment in other dimensions as a supplement to the global alignment is a solution.

Local alignment. As shown in Figure 3 (b), the regions or patches within an image and words in a sentence correspond with each other, so-called the local alignment. Global and local alignments form a complementary solution for ITR, which is a popular option and is classified as the local alignment-involved approach. Adopting the vanilla attention mechanism [Lee *et al.*, 2018; Wang *et al.*, 2019; Chen *et al.*, 2020c; Kim *et al.*, 2021] is a trivial way to explore the semantic region/patch-word correspondences. However, due to the semantic complexity, these approaches may not well catch the optimal fine-grained correspondences. For one thing, attending to local components selectively is a solution for searching for an optimal local alignment. [Liu *et al.*, 2019] made the first attempt to align the local semantics across modalities selectively. [Chen *et al.*, 2020a] and [Zhang *et al.*, 2020] were not far behind. The former learned to associate local components with an iterative local alignment scheme. And the latter noticed that an object or a word might have different semantics under the different global contexts and proposed to adaptively select informative local components based on the global context for the local alignment. After that, some approaches with the same goal as the above have been successively proposed with either designing an alignment guided masking strategy [Zhuge *et al.*, 2021] or developing an attention filtration technique [Diao *et al.*, 2021]. For another thing, achieving the local correspondence in a comprehensive manner is also a pathway to approximate an optimal local alignment. [Wu *et al.*, 2019a] enabled different levels of textual components to align with regions of images. [Ji *et al.*, 2021] proposed a step-wise hierarchical alignment network that achieves the local-to-local, global-to-local and global-to-global alignments.

Other than these, as shown in Figure 3 (c), there is another type of local alignment, *i.e.*, the relation-aware local alignment that can promote fine-grained alignment. These approaches [Xue *et al.*, 2021; Wei *et al.*, 2020] explore the intra-modal relation for facilitating inter-modal alignment. In addition, some approaches [Li *et al.*, 2019a; Yu *et al.*, 2021a; Diao *et al.*, 2021] model the image/text data as a graph structure with the edge conveying the relation information, and infer relation-aware similarities with both the local and global alignments by the graph convolutional network. Beyond these, [Ren *et al.*, 2021] considered the relation consistency, *i.e.*, the consistency between the visual relation among the objects and the textual relation among the words.

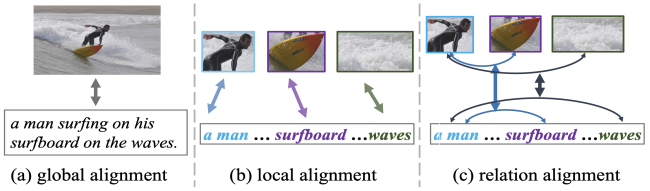

Figure 3: Illustration of different feature alignment architectures.

2.3 Retrieval Efficiency

Combining the feature extraction in Section 2.1 with the feature alignment in Section 2.2 makes up a complete ITR system with attention to the retrieval accuracy. Beyond that is the retrieval efficiency that is crucial for obtaining an excellent ITR system, thereby triggering a series of efficiency-focused ITR approaches.

Hash encoding. The hash binary code introduces the advantage on the model's computation and storage, lightening a growing concern over the hash encoding approaches for ITR. These studies learn to map the sample's feature to a compact hash code space to achieve a high-efficiency ITR. [Yang *et al.*, 2017] learned real-valued features and binary hash features of image and text simultaneously for benefiting from each other. [Zhang *et al.*, 2018] introduced an attention module to find the attended regions and words to prompt the binary feature learning. Besides these approaches in a supervised setting, unsupervised cross-modal hashing is also a concern. [Li *et al.*, 2018] incorporated the adversarial network into the unsupervised cross-modal hashing to maximize the semantic correlation and consistency between two modalities. [Yu *et al.*, 2021b] designed a graph-neighbor network to explore the sample's neighbor information for unsupervised hashing learning. The hash encoding approach benefits efficiency, yet also causes accuracy degradation due to the simplified feature representation with the binary code.

Model compression. With the advent of the cross-modal pre-training age, ITR takes a giant leap forward in accuracy at the expense of efficiency. The pre-training ITR approaches are usually characterized by the bulky network architecture, which gives birth to the model compression approach. Some researchers [Gan *et al.*, 2022] introduce the lottery ticket hypothesis to strive for smaller and lighter network architecture. Moreover, based on the consensus, i.e., the image preprocessing process takes the most significant computing resource consumption in the pre-training architecture, some researchers [Huang *et al.*, 2020; Huang *et al.*, 2021] specifically optimize the image preprocessing process to improve the retrieval efficiency. However, even with a lightweight archi-

tecture, most of these approaches that usually use the cross-attention for better feature learning still need to take a long reference time due to the quadratic executions of feature extraction.

Fast-then-slow. The above two groups cannot bring the best compromise between efficiency and accuracy, raising the third group: the fast-then-slow approach. Given that the VSE and CA approaches in Section 2.1 have the efficiency and accuracy advantages, respectively, several researchers [Miech *et al.*, 2021; Li *et al.*, 2021] propose first to screen out large amounts of easy-negative galleries by the fast VSE technology and then retrieve the positive galleries by the slow CA technology, thereby striving for a good balance between efficiency and accuracy.

3 Pre-training Image-text Retrieval

For the ITR task, the early paradigm is to fine-tune the networks that have been pre-trained on the computer vision and natural language processing domains, respectively. The turning point came in 2019, and there was an explosion of interest in developing a universal cross-modal pre-training modal and extending it to the downstream ITR task [Lu *et al.*, 2019; Li *et al.*, 2019b]. Under the powerful cross-modal pre-training technology, the ITR task experiences explosive growth in performance without any bells and whistles. Most of the pre-training ITR approaches currently adopt the transformer architecture as the building block. On this foundation, the research mainly focuses on model architecture, pre-training task and pre-training data.

Model architecture. A batch of works [Lu *et al.*, 2019; Li *et al.*, 2021] is interested in the two-stream model architecture, *i.e.*, two independent encodings followed by an optional later-interaction on the image and text processing branches. Meanwhile, the one-stream architecture encapsulating the image and text processing branches into one gains popularity [Li *et al.*, 2019b; Li *et al.*, 2020a; Kim *et al.*, 2021]. Most approaches heavily rely on the image preprocessing process that usually involves an object detection module or convolution architecture for extracting the preliminary visual features and as the input of the follow-up transformer. The resulting problems are twofold. Firstly, this process consumes more computational resources than the subsequent processes, leading to the model's inefficiency. Then, the predefined visual vocabulary from the object detection limits the model's expression ability, resulting in inferior accuracy.

Encouragingly, the research on improving the image preprocessing process has recently come into fashion. Regarding improving efficiency, [Huang *et al.*, 2021] adopted a fast visual dictionary to learn the whole image's feature. [Huang *et al.*, 2020] directly aligned the image pixels with the text in the transformer. Alternatively, [Kim *et al.*, 2021; Gao *et al.*, 2020] fed the patch-level features of the image into the transformer and [Liu *et al.*, 2021] segmented the image into grids for aligning with the text. In advancing accuracy, [Zhang *et al.*, 2021] developed an improved object detection model to promote visual features. [Xu *et al.*, 2021] put the tasks of object detection and image captioning together to enhance visual learning. [Xue *et al.*, 2021] explored the visual relation by adopting the self-attention mechanism when learning the image feature Taking all these into account, [Dou *et al.*, 2021] investigated these model designs thoroughly and presented an end-to-end new transformer framework, reaching a win-win between efficiency and accuracy. The advance of the cross-modal pre-training model architecture pushes forward the progress of ITR in performance.

Pre-training task. The pre-training pretext task guides the model to learn effective multimodal features in an end-to-end fashion. The pre-training model is designed for multiple cross-modal downstream tasks, hence various pretext tasks are usually invoked. These pretext tasks fall into two main categories: image-text matching and masked modeling.

The ITR is an important downstream task in the cross-modal pre-training domain and its associated pretext task, *i.e.*, the image-text matching, is well received in the pre-training model. In general, an ITR task-specific head is appended on the top of the transformer-like architecture to distinguish whether the input image-text pair is semantically matched by comparing the global features across modalities. It can be viewed as an image-text coarse-grained matching pretext task [Lu *et al.*, 2019; Li *et al.*, 2020a; Chen *et al.*, 2020c; Li *et al.*, 2021; Kim *et al.*, 2021]. Furthermore, it is also expanded to the image-text fine-grained matching pretext tasks: patch-word alignment [Kim *et al.*, 2021], region-word alignment [Chen *et al.*, 2020c] and region-phrase alignment [Liu *et al.*, 2021]. There is no doubt that the pre-training image-text matching pretext task establishes a direct link to the downstream ITR task, which narrows the gap between the task-agnostic pre-training model and ITR.

Inspired by the pre-training in the natural language processing, the masked language modeling pretext task is commonly used in the cross-modal pre-training model. Symmetrically, the masked vision modeling pretext task also emerges in this context. Both are collectively called the masked modeling task. In the masked language modeling task [Lu *et al.*, 2019; Li *et al.*, 2020a; Zhang *et al.*, 2021], the input text follows a specific masking rule that masks out several words in a sentence at random, and then this pretext task drives the network to predict the masked words based on the unmasked words and the input image. In the masked vision modeling task, the network regresses the masked region's embedding feature [Chen *et al.*, 2020c] or predicts its semantic label [Li *et al.*, 2020a] or does both [Liu *et al.*, 2021]. The masked modeling tasks implicitly capture the dependencies between the image and text, providing powerful support to the downstream ITR task.

Pre-training data. The research on the data level is an active trend in the cross-modal pre-training domain. For one thing, the intra- and cross-modal knowledge in the image and text data are fully exploited in the pre-training ITR approaches [Li *et al.*, 2020c; Cui *et al.*, 2021]. For another, many studies concentrate on increasing the scale of pre-training data. Beyond the most widely used large-scale out-of-domain datasets, especially for the pre-training model [Li *et al.*, 2020a; Li *et al.*, 2021], the in-domain datasets originally for fine-tuning and evaluating the downstream tasks are added into the pre-training data for better multimodal feature learning [Li *et al.*, 2020c; Li *et al.*, 2021]. Besides this, the

Type	Method	COCO1K		COCO5K		Flickr30K	
		t2i	i2t	t2i	i2t	t2i	i2t
VSE	VSE++ [Faghri *et al.*, 2017]	52.0	64.6	30.3	41.3	39.6	52.9
	LTBNN [Wang *et al.*, 2018]	43.3	54.9	-	-	31.7	43.2
	TIMAM [Sarafianos *et al.*, 2019]	-	-	-	-	42.6	53.1
	CVSE [Wang *et al.*, 2020]	59.9	74.8	-	-	52.9	73.5
	PCME [Chun *et al.*, 2021]	54.6	68.8	31.9	44.2	-	-
	ALIGN* [Jia *et al.*, 2021]	-	-	59.9	77.0	84.9	95.3
CA	SCAN [Lee *et al.*, 2018]	58.8	72.7	38.6	50.4	48.6	67.4
	SAN [Ji *et al.*, 2019]	69.1	85.4	46.2	65.4	60.1	75.5
	Vilbert* [Lu *et al.*, 2019]	-	-	-	-	58.2	-
	Oscar* [Li *et al.*, 2020c]	78.2	89.8	57.5	73.5	-	-
	Uniter* [Chen *et al.*, 2020c]	-	-	52.9	65.7	75.6	87.3
	Rosita* [Cui *et al.*, 2021]	-	-	54.4	71.3	74.1	88.9
	ALEBF* [Li *et al.*, 2021]	-	-	60.7	77.6	85.6	95.9
SA	DIME [Qu *et al.*, 2021]	64.8	78.8	43.1	59.3	63.6	81.0

Table 1: Accuracy comparison at R@1 among the ITR approaches from the perspective of the feature extraction. The approach marked with '$*$' represents the pre-training approach. We show the best results of each approach reported in the original paper.

rich non-paired single-modal data can be added into the pre-training data for learning more generalizable features [Li *et al.*, 2020b]. Other than all of these, some researchers [Qi *et al.*, 2020; Jia *et al.*, 2021; Yao *et al.*, 2022] collect new larger-scale data for the pre-training model, and such a simple and crude operation usually brings outstanding performance on various downstream cross-modal tasks, including ITR. In general, the focus at the data level positively affects the cross-modal pre-training model, naturally boosting the downstream ITR task.

4 Datasets and Evaluation

4.1 Datasets

The researchers have proposed various datasets for ITR. We summarize the most frequently used datasets as follows. 1) **COCO Captions** contains $123,287$ images collected from the Microsoft Common Objects in COntext (COCO) dataset, together with human generated five captions for each image. The average length of captions is 8.7 after a rare word removal. The dataset is split into $82,783$ training images, $5,000$ validation images and $5,000$ test images. The researchers evaluate their models on the 5 folds of 1K test images and the full 5K test images. 2) **Flickr30K** consists of $31,000$ images collected from the Flickr website. Each image contains five textual descriptions. The dataset is divided into three parts, $1,000$ images for validation, $1,000$ images for the test, and the rest for training.

4.2 Evaluation Metric

R@K is the most commonly used evaluation metric in ITR and is the abbreviation for recall at K-th in the ranking list, defined as the proportion of correct matchings in top-K retrieved results.

4.3 Accuracy Comparison

We compare the representative and latest ITR approaches in terms of accuracy from two perspectives: feature extraction and feature alignment.

Feature extraction. We present the comparison results in Table 1. For comparison among the VSE based approaches, ALIGN [Jia *et al.*, 2021] achieves substantial improvement over others on accuracy thanks to the large-scale pre-training on more than one billion image-text pairs that far surpass the amount of data from other pre-training approaches. For comparison among the CA approaches, we can see that the accuracy is improved gradually by these approaches over time. For comparison between the VSE based and CA approaches, 1) SCAN [Lee *et al.*, 2018], as the first attempt to the CA approach, makes a breakthrough at accuracy compared to the VSE based approach LTBNN [Wang *et al.*, 2018] at that time; 2) taken as a whole, the CA approaches have the overwhelming advantage over the VSE based ones at R@1 aside from ALIGN, which is attributed to the in-depth exploration of cross-modal feature interaction in the CA approaches. Nevertheless, as an exception, the VSE based approaches pre-training on extraordinarily massive data might offset the inferior performance caused by the less exploration on cross-modal interaction, which has been strongly supported by the results of ALIGN. For comparison between the SA approaches and the VSE based and CA ones, under the same setting, *i.e.*, conventional ITR, the SA approach DIME [Qu *et al.*, 2021] outperforms the VSE based and CA approaches on Flickr30k, and is inferior to the SAN [Ji *et al.*, 2019] on COCO Captions. There exists room for further development of the SA technology.

Type	Method	COCO1K		COCO5K		Flickr30K	
		t2i	i2t	t2i	i2t	t2i	i2t
Global.	VSE++ [Faghri *et al.*, 2017]	52.0	64.6	30.3	41.3	39.6	52.9
	LTBNN [Wang *et al.*, 2018]	43.3	54.9	-	-	31.7	43.2
	SEAM [Wu *et al.*, 2019b]	57.8	71.2	-	-	52.4	69.1
	TIMAM [Sarafianos *et al.*, 2019]	-	-	-	-	42.6	53.1
	Dual-path [Zheng *et al.*, 2020]	47.1	65.6	25.3	41.2	39.1	55.6
	PCME [Chun *et al.*, 2021]	54.6	68.8	31.9	44.2	-	-
	ALIGN* [Jia *et al.*, 2021]	-	-	59.9	77.0	84.9	95.3
Local.	SCAN [Lee *et al.*, 2018]	58.8	72.7	38.6	50.4	48.6	67.4
	CAMP [Wang *et al.*, 2019]	58.5	72.3	39.0	50.1	51.5	68.1
	VSRN [Li *et al.*, 2019a]	62.8	76.2	40.5	53.0	54.7	71.3
	IMRAM [Chen *et al.*, 2020a]	61.7	76.7	39.7	53.7	53.9	74.1
	MMCA [Wei *et al.*, 2020]	61.6	74.8	38.7	54.0	54.8	74.2
	Uniter* [Chen *et al.*, 2020c]	-	-	52.9	65.7	75.6	87.3
	SGRAF [Diao *et al.*, 2021]	63.2	79.6	41.9	57.8	58.5	77.8
	SHAN [Ji *et al.*, 2021]	62.6	76.8	-	-	55.3	74.6
	ViLT* [Kim *et al.*, 2021]	-	-	42.7	61.5	64.4	83.5
	ALBEF* [Li *et al.*, 2021]	-	-	60.7	77.6	85.6	95.9

Table 2: Accuracy comparison at R@1 among the ITR approaches from the perspective of the feature alignment. The approach marked with '$*$' represents the pre-training approach. Global. and Local. are short for the global-driven alignment and local-involved alignment approaches, respectively. We show the best results of each approach reported in the original paper.

Feature alignment. The comparison results are shown in Table 2. In terms of comparison within the global alignment-driven approaches, even with a basic two-stream architecture for the global alignment, ALIGN is still on top of other approaches at R@1, including TIMAM [Sarafianos *et al.*, 2019] and PCME [Chun *et al.*, 2021] with the sophisticated network architecture for the global alignment. In terms of comparison within the local alignment-involved approaches, ALBEF [Li *et al.*, 2021] displays excellent performance. It is worth noting that Uniter [Chen *et al.*, 2020c] and ViLT [Kim *et al.*, 2021] only with the vanilla attention mechanism can

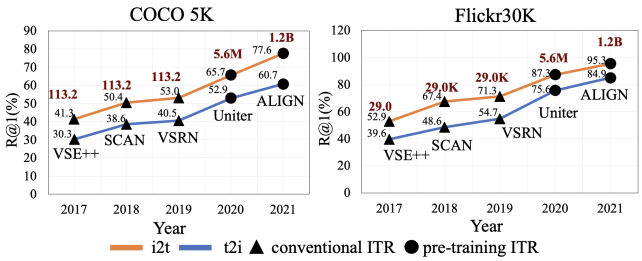

Figure 4: The development trend of ITR in recent years. The number with black above the line chart is the R@1 value of the approach, and the number with red is the amount of the multimodal training data.

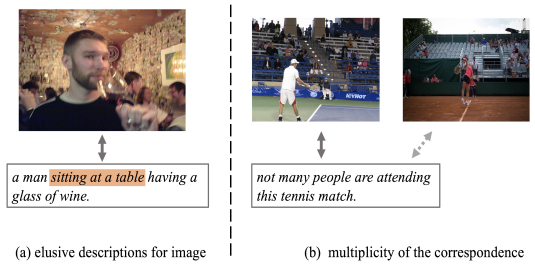

Figure 5: Illustration of noise data in the COCO Captions. (a) It is difficult to capture the content in the image based on the paired textual description highlighted by orange. (b) In addition to the positive image-text pair with a solid arrow, it seems to be correspondence for the negative image-text pair with a dotted arrow.

get decent results. By contrast, SCAN [Lee *et al.*, 2018] and CAMP [Wang *et al.*, 2019] with a similar mechanism are underwhelming at R@1. The Uniter and ViLT conduct the ITR task in a pretrain-then-finetune form, and the rich knowledge from the pre-training cross-modal data benefits the downstream ITR task. In terms of comparison between the global alignment-driven and local alignment-involved approaches, the latter shows better performance than the former on the whole, indicating the importance of local alignment for achieving high-accuracy ITR.

Furthermore, we summarize the development trend of ITR from 2017 to 2021 in Figure 4. A clear trend of increasing accuracy can be seen over the years. Specifically, the big jump comes in 2020 thanks to the pre-training ITR technology. After this, the accuracy of the pre-training ITR approach continues to keep the momentum of development. It follows that the pre-training ITR technology plays a leading role in promoting ITR development. It can not be separated from the support of the enlarging scale of training data. We can observe a dramatic increase in the amount of training data with the coming of the pre-training ITR.

5 Conclusion and Future Works

In this paper, we presented a comprehensive review of ITR approaches from four perspectives: feature extraction, feature alignment, system efficiency and pre-training paradigm. We also summarized extensively used datasets and evaluation metric in ITR, based on which we quantitatively analyzed the performance of the representative approaches. It concludes that ITR technology has made considerable development over the past few years, especially with the coming of the cross-modal pre-training age. However, there still exist some less-explored issues in ITR. We make some interesting observations on possible future developments as follows.

Data. The current ITR approaches are essentially data-driven. In other words, the researchers design and optimize the network for seeking an optimal retrieval solution based on available benchmark datasets. For one thing, the heterogeneity and semantic ambiguity of cross-modal data can inevitably introduce noise into the datasets. For example, as shown in Figure 5, there exist the elusive textual description for the image and the multiplicity of the correspondences between the images and texts in the COCO Captions. To some extent, therefore, the results of current ITR approaches on such datasets remain controversial. There have been a few explorations about the data multiplicity [Song and Soleymani, 2019; Chun *et al.*, 2021; Hu *et al.*, 2021], yet only considering the training data and ignoring the test one. For another thing, beyond the vanilla data information, *i.e.*, the image and text, the scene-text appearing in images is a valuable clue for ITR, which is usually ignored in the existing approaches. [Mafla *et al.*, 2021] is a pioneer work to explicitly incorporate the scene-text information into ITR model. These studies leave room for further ITR development at the data level.

Knowledge. Humans have the powerful ability to establish semantic connections between vision and language. It benefits from their cumulative commonsense knowledge, together with the capacity of causal reasoning. Naturally, incorporating this high-level knowledge into the ITR model is valuable for improving its performance. CVSE [Wang *et al.*, 2020] is a pioneer work that computes the statistical correlations in the image captioning corpus as the commonsense knowledge for ITR. However, such commonsense knowledge is constrained by the corpus and is not a perfect fit for ITR. It might be promising to tailor-make a commonsense knowledge and model the causal reasoning for ITR in the future.

New paradigm. Under the current trend, the pre-training ITR approaches have an overwhelming advantage on accuracy compared to the conventional ITR ones. The pretrain-then-finetune over a large-scale cross-modal model becomes a fundamental paradigm for achieving state-of-the-art retrieval results. However, this paradigm with the need for large amounts of labeled data in the finetun phase is hard to apply in real-world scenarios. It is meaningful to seek and develop a new resource-friendly ITR paradigm. For example, the recently budding prompt-based tuning technology with an excellent few-shot capability provides a guide for developing such a new paradigm, so-called pretrain-then-prompt.

Acknowledgements

This work is supported by the National Science Foundation of China under Grant NSFC 62002252, and is also partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization, and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions. Juntao Li is the corresponding author of this paper.

References

[Baltrušaitis *et al.*, 2018] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 2018.

[Chen *et al.*, 2020a] 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 *CVPR*, 2020.

[Chen *et al.*, 2020b] Jianan Chen, Lu Zhang, Cong Bai, and Kidiyo Kpalma. Review of recent deep learning based methods for image-text retrieval. In *MIPR*, 2020.

[Chen *et al.*, 2020c] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

[Chun *et al.*, 2021] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, 2021.

[Cui *et al.*, 2021] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In *ACM MM*, 2021.

[Diao *et al.*, 2021] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, 2021.

[Dou *et al.*, 2021] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. An empirical study of training end-to-end vision-and-language transformers. *arXiv preprint arXiv:2111.02387*, 2021.

[Faghri *et al.*, 2017] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2017.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013.

[Gan *et al.*, 2022] Zhe Gan, Yen-Chun Chen, Linjie Li, Tianlong Chen, Yu Cheng, Shuohang Wang, and Jingjing Liu. Playing lottery tickets with vision and language. In *AAAI*, 2022.

[Gao *et al.*, 2020] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *SIGIR*, 2020.

[Hu *et al.*, 2021] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *CVPR*, 2021.

[Huang *et al.*, 2020] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[Huang *et al.*, 2021] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, 2021.

[Huo *et al.*, 2021] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021.

[Ji *et al.*, 2019] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. Saliency-guided attention network for image-sentence matching. In *ICCV*, 2019.

[Ji *et al.*, 2021] Zhong Ji, Kexin Chen, and Haoran Wang. Step-wise hierarchical alignment network for image-text matching. In *IJCAI*, 2021.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.

[Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.

[Li *et al.*, 2018] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *CVPR*, 2018.

[Li *et al.*, 2019a] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019.

[Li *et al.*, 2019b] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[Li *et al.*, 2020a] 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 *AAAI*, 2020.

[Li *et al.*, 2020b] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[Li *et al.*, 2020c] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[Li *et al.*, 2021] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision

and language representation learning with momentum distillation. In *NeurIPS*, 2021.

[Liu *et al.*, 2019] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *ACM MM*, 2019.

[Liu *et al.*, 2021] Yongfei Liu, Chenfei Wu, Shao-yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. In *EMNLP*, 2021.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[Mafla *et al.*, 2021] Andrés Mafla, Rafael S Rezende, Lluis Gomez, Diane Larlus, and Dimosthenis Karatzas. Stacmr: scene-text aware cross-modal retrieval. In *WACV*, pages 2220–2230, 2021.

[Miech *et al.*, 2021] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, 2021.

[Qi *et al.*, 2020] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

[Qu *et al.*, 2021] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. Dynamic modality interaction modeling for image-text retrieval. In *SIGIR*, 2021.

[Ren *et al.*, 2021] Shuhuai Ren, Junyang Lin, Guangxiang Zhao, Rui Men, An Yang, Jingren Zhou, Xu Sun, and Hongxia Yang. Learning relation alignment for calibrated cross-modal retrieval. In *ACL-IJCNLP*, 2021.

[Sarafianos *et al.*, 2019] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *ICCV*, 2019.

[Song and Soleymani, 2019] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, 2019.

[Uppal *et al.*, 2022] Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. Multimodal research in vision and language: A review of current and emerging trends. In *Information Fusion*, 2022.

[Wang *et al.*, 2018] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. In *TPAMI*, 2018.

[Wang *et al.*, 2019] 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 *ICCV*, 2019.

[Wang *et al.*, 2020] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, 2020.

[Wei *et al.*, 2020] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *CVPR*, 2020.

[Wu *et al.*, 2019a] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *CVPR*, 2019.

[Wu *et al.*, 2019b] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *ACM MM*, 2019.

[Xu *et al.*, 2021] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. In *ACL-IJCNLP*, 2021.

[Xue *et al.*, 2021] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. In *NeurIPS*, 2021.

[Yang *et al.*, 2017] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, 2017.

[Yao *et al.*, 2022] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2022.

[Yu *et al.*, 2021a] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. In *AAAI*, 2021.

[Yu *et al.*, 2021b] Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *AAAI*, 2021.

[Zhang *et al.*, 2018] Xi Zhang, Hanjiang Lai, and Jiashi Feng. Attention-aware deep adversarial hashing for cross-modal retrieval. In *ECCV*, 2018.

[Zhang *et al.*, 2020] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *CVPR*, 2020.

[Zhang *et al.*, 2021] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021.

[Zheng *et al.*, 2020] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. In *TOMM*, 2020.

[Zhuge *et al.*, 2021] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *CVPR*, 2021.