



Disentangled OCR: A More Granular Information for “Text”-to-Image Retrieval

Xinyu Zhou, Shilin Li, Huen Chen, and Anna Zhu^(✉)

Wuhan University of Technology, Wuhan, China
{297932,shilinli,259776,annazhu}@whut.edu.cn

Abstract. Most of the previous text-to-image retrieval methods were based on the semantic matching between text and image locally or globally. However, they ignore a very important element in both text and image, i.e., the OCR information. In this paper, we present a novel approach to disentangle the OCR from both text and image, and use the disentangled information from the two different modalities for matching. The matching score is consist of two parts, the traditional global semantic text-to-image representation matching and OCR matching scores. Since there is no dataset to support the training of text OCR disentangled task, we label partial useful data from TextCaps dataset, which contains scene text images and their corresponding captions. We relabel the text of captions to OCR and non-OCR words. In total, we extract 110K captions and 22K images from TextCaps, which contain OCR information. We call this dataset TextCaps-OCR. The experiments on TextCaps-OCR and another public dataset CTC (COCO-Text Captions) demonstrate the effectiveness of disentangling OCR in text and image for cross modality retrieval task.

Keywords: Text-to-image retrieval · OCR · Disentangled information · Cross modality

1 Introduction

Text-image cross-modal retrieval has attracted wide attention and made great progress in recent years. It aims to return the most relevant images to the query text. Its performance mainly relies on the consistency and alignment of visual and language representations.

Most of the previous image retrieval tasks were based on the similarity of textual and visual features. With the advent of Transformer [2] in recent years that induces cross attention for contextual feature representations, researchers start to use it [2] for highly coupled training based on visual objects in images, such as SCAN and X-VLM [3,4], and for matching through scene graph, such as LGSGM and SGRAF [5,6]. They consider the granular information of visual objects in images but ignore very important information, namely OCR.



Query: A soccer player wearing a yellow jersey with the number 3.

Fig. 1. Defects in current retrieval model, which cannot combining OCR and image semantic information.

Given a query text as shown in Fig. 1, the general text-to-image retrieval model may get the image retrieval results without scene text as Fig. 1(b), which is caused by unable to recognize the OCR information “3”. But the missing information “3” is the most important element of the query sentence. If we extract the OCR number “3” from the query and only use it for retrieval like JTSL [23], it may output the image with number “3” but has no relationship with the semantic description of the text query. Obviously these conventional methods are not feasible to text query containing OCR information. It inspires us to combine the two strategies above to perform for more granular OCR-based retrieval. We call this task as “Text”-to-Image retrieval, where “Text” represents linguistic description and its OCR information corresponding to scene text of image.

There are some related works [15, 16] mentioning this point, but they only match query with the OCR information in images globally. But, it leads to increasing errors with the image amount growing, because the scene text in images may contain words in query. So we distinguish the OCR information in both query text and images, then perform matching on two aspects, one for the disentangled OCR and the other for the original semantic matching on the two modalities. Therefore, we call this process as OCR disentangling. We refer to this kind of OCR as text OCR in later sections.

However, OCR disentangling is not as simple as expected. Since OCR in images, i.e., scene text, are easy to be detected as a special visual object, but OCR in query texts cannot be well distinguished by named entity recognition methods or POS tagging. Text OCR can be any component of a sentence requiring some reasoning to be detected. Humans rely more on context to infer text OCR, but sometimes they cannot extract the OCR information completely and correctly. Therefore, we can complete this task through the model that can obtain context information, but the most critical issue is the lack of text OCR annotation datasets.

Therefore, we put forward a dataset called TextCaps-OCR for text OCR detection, and we use Bert [9] to solve this new task. Moreover, we establish one model containing two retrieval ways. The first is a word set based approach. We transforms images into captions using image captioning model, then separately

split the query texts and captions to get two kinds of sets, finally we perform retrieval through the similarity between the sets from query and image. The second is based on image and query features' similarity. Both ways use the disentangled OCR for matching.

The comparative experiments are carried out on TextCaps-OCR and CTC (COCO-Text Caps) [15], and the results illustrate the correctness of our motivation and the effectiveness of this method.

Thus, we mainly have the following **three contributions**:

1. With strong motivation, we introduce OCR as a more granular information into the text-to-image retrieval task and demonstrate its feasibility and effectiveness.
2. We provide a new dataset called TextCaps-OCR with text OCR annotation, and propose a simple retrieval method to match both global semantic and local OCR in different modalities. Additionally, we design a Soft Matching strategy to boost the robustness of OCR matching in the model.
3. Experimental results on TextCaps-OCR and CTC [15] (COCO Text Caps) prove the effectiveness of the OCR disentangling operation. Our method outperforms the X-VLM model [4] (the SOTA method on Flickr30K and COCO) on TextCaps-OCR and achieve comparable results with a great pre-training model on CTC.

2 Related Work

2.1 OCR Related Cross-Modal Tasks

OCR information has been used in VQA (Visual Question Answering), image captioning, and other areas. Since the answers to VQA are usually present in the image, it is necessary to extract and add the OCR information to the answer lexicon. As for image captioning, to describe an image by text, considering OCR information in the image will greatly enrich its granularity of description. M4C [7] is the first model using OCR information for VQA and Image Caption tasks. It consists of three parts. Firstly, they extract the object features by Faster-RCNN [24], and then use Sence Text detection network to obtain the representations of OCR in the image. Secondly, they pass the obtained information and query into a large Transformer [2] encoder after embedding them separately to learn the relationship between the features of different modalities. Lastly, they add the OCR to the answer dictionary through pointer networks [25] and output the most likely answer through the decoder. Another effective VQA model is called SBD [26], firstly splitting the text features into two functionally distinct parts, a linguistic part and a visual part, which flow into the corresponding attention branches. The encoded features are then fed into the Transformer's [2] decoder to generate answers or captions, using three attention blocks to filter out irrelevant or redundant features and aggregate them into six separate functional vectors. And they demonstrate that OCR is the main contributor to VQA, while visual semantics only plays a supporting role. And with the introduction of M4C [7],

M4C-Captioner [1] followed, which has the same architecture as M4C [7], but is applied to the task of image captioning by removing the question input and directly use its multi-word answer decoder to generate captions.

Through our investigation, we find two existing studies on retrieval using OCR. The first is StacMR [15]. They mainly propose a dataset called CTC (COCO Text Captions) and conduct several comparative experiments based on existing models [3, 10, 21] to prove the effectiveness of OCR in retrieval tasks. The second is ViSTR [16]. They use vision and OCR aggregated transformer, i.e. cross-modal learning of images, text and scene text through separate transformers. But they both perform retrieval matching the features globally, which we have previously shown to be flawed.

2.2 Cross-Modal Retrieval

There are two main types of cross-modal retrieval tasks, one is to extract the corresponding modal features separately and then perform similarity matching directly, and the second is to take multi-modal information together as input and perform similarity output by cross-attention. The former takes much less time than the latter, but the effect is relatively poor, so we can call the former Fast and the latter Slow. The most typical Fast model is SCAN [3]. It first extracts the objects in images through Bottom-up attention [27], and then compute the similarity among all of the object and word features by cosine similarity to rank. And surprisingly there are also large-scale pre-training models in Fast, represented by CLIP and BLIP [11, 12]. They are both trained by using the similarity of positive samples as the numerator and negative samples as the denominator in the loss function. Their high accuracy for retrieval is supported more by computing power and the huge training data. Facing OCR-related text-to-image retrieval, they fail in many cases. And the model of the Slow class is represented by Pixel Bert [28], which imitates the Bert [2] by randomly sampling pixel-level image information, concating it with word features as input, and training it through MLM. Subsequently, there are many scene graph-based retrieval methods, such as LGSGM and SGRAF [5, 6], both of them retrieve by matching the semantics of objects and their relationships. What deserves a special mention is that the research “Thinking Fast and Slow” [29] bridges the Fast and Slow models through distillation learning. They use the Slow model as a teacher to do distillation learning on the Fast model, generating the model-Fast&Slow, and then performing the whole retrieval through Fast&Slow to get top-20 returns. Finally, they re-rank the top-20 images through the slow model to get the final result. But unfortunately, none of these approaches pay attention to the importance of OCR.

3 Methods

This section mainly describes the methods for text OCR detection and two embarrassingly simple but effective retrieval ways combined with disentangled OCR information. The model framework is shown in Fig. 2.

Input Text (x):

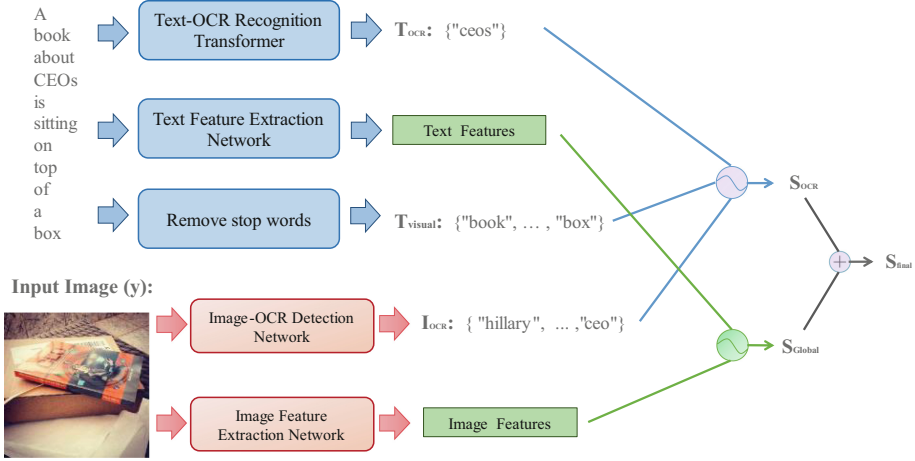


Fig. 2. Our retrieval model can be divided into two parts, the first part is to compute the score of OCR using Soft Matching (blue part), the second part is the to compute the similarity score between text x and image y using the cosine similarity which can be shown as $f(x)@g(y)$ (green part). (Color figure online)

3.1 Text OCR Detection

We use Bert [9] to perform **TOD** (**T**ext **O**CR **D**etection). Since TOD is a task that heavily relies on text comprehension, we directly mark OCR words as 1 and background (non-OCR) words as 0 to fit the training way of Bert, i.e. MLM. What deserves a special mention is that we think this decoupling method is also one of the most important reasons why we can achieve such good results.

We use the cross-entropy of the last hidden layer and the real label as the loss. And the weight of the loss function is (5/6, 1/6), because the ratio of OCR to background words (non OCR) is close to 1:5. Considering the last n -dimensional hidden layer: $H = [H_1, H_2, \dots, H_n]$, and the real label: $G_t = [G_{t_1}, G_{t_2}, \dots, G_{t_n}]$, then we can calculate the loss as:

$$L_{CE} = -\frac{1}{n} \sum_{i=0}^n [G_{t_i} * \ln(H_i) + (1 - G_{t_i}) * \ln(1 - H_i)] \quad (1)$$

3.2 Methods to Perform Retrieval with Disentangled OCR

In order to prove the effectiveness of OCR and OCR decoupling in different modalities, we use a simple and easy method: 1. Text-to-Image retrieval based on set-to-set matching. 2. Image and text feature matching through cosine similarity. It is acknowledged that the effect of fast model is poor. To better compute the matching degree, we propose a soft matching strategy, which means conducting the first match through global OCR matching, i.e., Scene Text (OCR in the

image) matching with all the words in a query, and then conducting the second match plus points through Text OCR, which can effectively increase the robustness of the model and method. This can be seen in the subsequent experimental results.

Soft Matching Introduction. The Soft Matching process can be simplified into pseudo code as shown in Algorithm 1. We use idf (inverse document frequency) to weight because there are prediction errors when detecting text OCR, and most of the words with prediction errors are common words in images, which makes the subsequent results worse. So, we introduce the idf weighting in information processing into our method, which can effectively suppress the impact of OCR prediction errors and greatly increase the robustness of the method. At last, we add the OCR score and set-to-set score directly to get the final score:

$$S_{final} = Softmax(S_{Global}) + S_{OCR} \quad (2)$$

Algorithm 1. Soft Matching

Input: the query *Text*, OCR in matching image *Scence_Text*, OCR in matching text *Text_OCR*, the times of every word appears in images *OCR_DICT*, the average times of words in images *OCR_DICT_AVG*. *i* refers to the *i*-th query, *j* refers to the *j*-th matching image/text and *n* refers to the number of images.

```

1: for word in Texti do
2:   if word in Scence_Textj then
3:     if word in Text_OCRj then
4:       SOCRj += 2/OCR_DICT[word]
5:     else
6:       SOCRj += 1/OCR_DICT_AVG
7:     end if
8:   end if
9: end for
10: SOCR = [SOCR1, ..., SOCRj, ..., SOCRn]
11: SOCR = softmax(SOCR)
Output: SOCR

```

Retrieval Based on Captions. Because there is no other image retrieval model using OCR before, we just use the existing image-captioning model to prove the potential of our approach. We use a pre-training model called OFA [20] to generate the captions on TextCaps-OCR-1K directly. After that, segment the sentence and remove the stop words. Then we can transform the feature matching into a set-to-set (sts) matching, which divides the intersection by the length of the query set as the global score, and the *i*-th score is calculated as:

$$S_{Global_{sts}i} = L(Query \cap Captions_i) / L(Query) \quad (3)$$

where L means the length of the corresponding set. At the same time, we can compute the OCR score through Soft Matching.

Retrieval Based on Features. The whole process is shown in Fig. 2. The difference from retrieval based on captions is that there are no words. We just compute the global score using the features from the extraction models by cosine similarity. For the text features $f(x)$ and image feature $g(x)$ (both are n -dimensions), the global score can be calculated as:

$$S_{Global_i} = \frac{\sum_{i=1}^n f(x)_i * g(x)_i}{\sqrt{\sum_{i=1}^n (f(x)_i)^2} * \sqrt{\sum_{i=1}^n (g(x)_i)^2}} \quad (4)$$

The final score is the same as Eq. 2. And we use the NCE Loss:

$$L_{NCE} = - \sum_{i=1}^n \ln \left(\frac{e^{f(x_i)^T g(y_i)}}{e^{f(x_i)^T g(y_i)} + \sum_{(x', y') \in X_i} e^{f(x')^T g(y')}} \right) \quad (5)$$

which contrasts the score of the positive pair (x_i, y_i) to a set of negative pairs sampled from a negative set X_i . Above all, the total loss can be calculated as:

$$L_{total} = L_{CE} + L_{NCE} \quad (6)$$

4 The TextCaps-OCR Dataset

This section mainly describes the proposed dataset: **TextCaps-OCR**. We first explained its origin, the form of data storage and how it is tailored for text OCR detection. Then we compared it with other conventional datasets.

```
"0c0f40dde4a4c770": {
  "captions": [
    "Laneige sleeping mask comes in a 2.7 fl oz size, and is packaged in a pink box.",
    "A box of Laneige Water Sleeping Pack is on a red mat.",
    "A package by Laneige of their water sleeping pack.",
    "The box contains a water sleeping pack made by the company Laneige.",
    "A box containing the water sleeping pack by Laneige"
  ],
  "caption_tokens": [
    ["Laneige", "sleeping", "mask", "comes", "in", "a", "2.7", "fl", "oz", "size", "and", "is", "packaged", "in", "a", "pink", "box"],
    ["A", "box", "of", "Laneige", "Water", "Sleeping", "Pack", "is", "on", "a", "red", "mat"],
    ["A", "package", "by", "Laneige", "of", "their", "water", "sleeping", "pack"],
    ["The", "box", "contains", "a", "water", "sleeping", "pack", "made", "by", "the", "company", "Laneige"],
    ["A", "box", "containing", "the", "water", "sleeping", "pack", "by", "Laneige"]
  ],
  "caption_ocr": [
    [1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
    [0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0],
    [0, 0, 0, 1, 0, 0, 1, 1, 1],
    [0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1],
    [0, 0, 0, 0, 1, 1, 1, 0, 1, 1]
  ],
  "ocr": ["laneige", "water", "sleeping", "laniege", "pack"]
}
```



Fig. 3. TextCaps-OCR's text data storage structure and its corresponding image.

4.1 Origin and Storage Structure

The **TextCaps-OCR** is a new dataset which contains labeled text OCR. We selected 21873 pictures with clear OCR from the TextCaps [1] for human annotation of the text OCR, and generated the OCR annotation corresponding to each caption, which is divided into 19130 training sets and 2743 test sets, in which each picture has 5 captions, and its storage form is as follows: Fig. 3. We can see that there is OCR annotation in the captions in TextCaps-OCR Dataset, which is very different from the previous dataset, in order to serve our subsequent task: Text OCR Detection. Text OCR Detection is a new task proposed by us to detect the OCR in text. We labeled the text OCR in captions in 1, and the background words should be 0, which can be seen in Fig. 3.

4.2 Comparison with Other Datasets

The most common retrieval datasets are Flickr30K, MS-COCO, COCO-Text and CTC [15, 17–19]. The contained OCR information comparison between these datasets and our newly proposed dataset TextCaps-OCR is shown in the Table 1.

Table 1. Comparison between TextCaps-OCR and other commonly datasets. Besides, \star is Scene Text, \dagger is Text OCR and \ddagger is Labeled Text OCR.

Dataset	The number of		Annotations		
	Images	Texts	ST^\star	TO^\dagger	LTO^\ddagger
Flicer30K [17]	31784	31784*5	✗	✗	✗
MS-COCO [18]	328K	328K*5	✗	✗	✗
COCO-Text [19]	29210	29210*5	✓	✗	✗
COCO-Text Caps [15]	10683	10683*5	✓	✓	✗
TextCaps [1]	28408	28408*5	✓	✓	✗
TextCaps-OCR [ours]	21873	21873*5	✓	✓	✓

The main difference between our dataset TextCaps-OCR and other datasets is the annotated text OCR, which mainly serves our task of text OCR detection. It is worth noting that the TextCaps [1] is mainly based on VQA and image captioning. However, through our investigation, we found that this dataset is also challenging for the retrieval task because there are many similar images. The difference lies in the description of details and OCR, so we specially selected these challenging images to enter our new dataset, which is very suitable for fine-grained retrieval tasks and models with an OCR recognition function.

5 Experiment

In this section, we evaluate the benefits of our methods. We evaluate the Text OCR Detection on TextCaps-OCR in Sect. 5.1. In Sect. 5.2, we analyze the

results of ablation experiments and qualitative comparisons. Moreover, we compare our approach with other published state-of-the-art retrieval methods on TextCaps-OCR and CTC [15] in Sect. 5.3.

5.1 Performance of Text OCR Detection

For training the Bert-base [9] for text OCR detection, we use the AdamW optimizer with a learning rate of $5e-5$ for just 2 epochs. It gets high detection performance as shown in Table 2. The precision and recall for text OCR are **91.1%** and **92.2%**, for background words are **97.4%** and **97.0%**. The high performance ensures the following process by using disentangled information for matching.

Table 2. Text OCR detection performance (%)

Class	Precision	Recall
Text OCR	91.1	92.2
Non-OCR words	97.4	97.0

5.2 Ablations

Firstly, we should mention the Scene Text Detection and Recognition model we used: Mask textspotter v3 [22]. Despite the fact that it incorporates OCR detection and identification, its effect is significantly inferior to that of today’s SOTA due to its age, so we also make comparative tests using the real value.

Table 3. Ablation study on the impact of various matching strategies



Model	TextCaps-OCR-1K		
	R@1	R@5	R@10
General matching	17.34	35.78	46.1
OCR + General matching	19.86	47.08	62.06
Disentangled OCR + General matching	57.84	71.36	78.58
Soft matching + General matching	58.56	72.36	78.84
OFA [20]	34.12	52.08	61.54
OCR + OFA	40.80	60.12	69.60
Disentangled OCR + OFA	65.16	80.80	85.54
Soft matching + OFA	65.98	81.08	85.58
Soft matching + OFA(tt)	75.22	87.14	90.74
Soft matching + OFA(ti)	76.1	87.66	91.44

Method Description. We conduct two categories of comparative experiments on TextCaps-OCR to prove the effectiveness of our methods: “General feature matching series” and “Captions-based set-to-set matching series”. For our approach, we consider a total of four modules: **1. General matching**, which does not use the OCR information. **2. OCR**, which means using the entangled OCR information. **3. Disentangled OCR**, i.e., using the disentangled OCR information. **4. Soft matching**, a matching way proposed by us, aims to match by combining both entangled OCR and disentangled OCR. Thus, the general feature matching series includes the following four ways: 1, 1 + 2, 1 + 3 and 1 + 4. As for the captions series, we removed OCR words from captions generated by OFA [20] to replace the basic model (i.e. general matching), and the rest is the same as above. For general matching, we use resnet50 to extract image features and Bert-base to extract text features, and MLP is added behind both networks (the former is 1000-500-100, the latter is 768-500-100, both containing the Dropout (0.9) and activation function Relu). Besides, the learning rate is $2e-4$ for 15 epochs.

Results Analysis. As shown in Table 3, compared with general matching, we can see that the entangled OCR information does not significantly improve the results, indicating the entangled OCR information is indeed limited. But on the contrary, the disentangled OCR information has made a qualitative leap in performance, improving R@1 by **40.5%**, R@5 by **35.58%** and R@10 by **32.74%**. Soft matching slightly improves the network performance again. And when we use a better basic network: OFA [20], the disentangled OCR information still makes a leap in the overall performance; relatively simple networks increase by an average of **10%**. Considering the error of OCR recognition (both text and image), we use the real value (**tt** refers to true text OCR and **ti** refers to true image OCR (i.e. Scene Text)). We can see that one of the true values of OCR can greatly improve the performance again. These three performance overflights can obviously prove the superiority of disentangled OCR on retrieval tasks.

Qualitative Comparisons. We also report some examples to illustrate the effectiveness and necessity of our method. As shown in Fig. 4, we **bold** the text OCR identified by our method in **two** text queries, which are obviously correct. Besides, for the first text query, the correct image could not be returned first using the entangled OCR. For the second text query, compared with using disentangled OCR to perform retrieval, entangled OCR did not return the correct image in the first three returned images. After investigation, we find that it is the tenth one that corresponds to the correct image. Meanwhile, we try to increase the weight of the OCR entangling matching score but ranking further down, which indicates that the robustness of entangled OCR is extremely poor when there are fewer OCR in the search sentences and more OCR information in images.

1. Text Query: Wonderful bottle of **Spring Seed Wine** from the country of **Australia**.

	(a)Top-1 ✓	(b)Top-2 ✗	(c)Top-3 ✗	(d)Top-1 ✗	(e)Top-2 ✓	(f)Top-3 ✗
Score:	2.3	2.0	1.5	2.0	1.6	1.5
						

2. Text Query: A large red and white commercial jet taxis on runway **101** right.



	(a)Top-1 ✓	(b)Top-2 ✗	(c)Top-3 ✗	(d)Top-1 ✗	(e)Top-2 ✗	(f)Top-3 ✗
Score:	2.2	2.0	1.5	1.505	1.504	1.503
						

Fig. 4. Examples of the text-to-image retrieval task for comparisons between results with disentangled and entangled OCR, where (a)(b)(c) are the results returned by decoupled OCR and (d)(e)(f) are the results returned by coupled OCR. In text queries, we use **bold** to mark the text OCR identified by our method.

5.3 Comparison to the State of the Art

The experimental results on TextCaps-OCR and CTC are shown in Table 4 and Table 5 respectively.

Performance on TextCaps-OCR. Considering X-VLM [4] (the SOTA method on both MS-COCO [18] and Flickr30K [17]) is a pre-training model containing 4M images, we do the zero-shot testing on our test dataset. Although X-VLM [4] has better performance than general matching and OFA (both without OCR information), our models easily outperform it through Soft Matching by at least **13.16%** at Recall@1. Unfortunately, we cannot compare our method with the ViSTA [16] because it’s not open source yet.

Table 4. Comparisons on TextCaps-OCR.

Model	TextCaps-OCR-1K		
	R@1	R@5	R@10
X-VLM [4]	45.4	70.38	78.56
General matching	17.34	35.78	46.1
Soft Matching + General matching	58.56	72.36	78.84
OFA [20]	34.12	52.08	61.54
Soft Matching + OFA	65.98	81.08	85.58

Table 5. Comparisons with the state-of-the-art scene text aware approaches on CTC.

Model	CTC-1K		
	R@1	R@5	R@10
SCAN [3]	26.6	53.6	65.3
VSRN [10]	26.6	54.2	66.2
STARNet [21]	31.5	60.8	72.4
ViSTA-S [16]	36.7	66.2	77.8
General matching	14.6	30.8	41.3
OCR + General matching	32.8	60.6	74.2
Disentangled OCR + General matching	26.2	52.7	62.8
Soft matching + General matching	35.4	68.0	74.9

Performance on CTC. We cannot use OFA [20] or X-VLM [4] on CTC because they are pre-trained on MS-COCO [18] which is the superset of CTC [15]. But, we still go beyond the traditional methods in all aspects through Soft matching + General matching. As for the pre-training model using the coupled OCR information, i.e. ViSTA [16], we can only achieve considerable results because of the problems with the dataset (CTC), the training scale (we just train on CTC, which only contains 9K images) and our too simple basic model. But even though there remain so many problems, we still checked only 1.3% on Recall@1 with the state of the art, and **1.8%** higher on Recall@5. Besides, combining the results of **OCR + General matching** and **Disentangled OCR + General matching** on these two datasets, it is not difficult to find that the OCR decoupling approach has considerably better generalization and performance than the coupling one, and Soft Matching does greatly improve the robustness of OCR matching and achieves the best results on both datasets.

6 Conclusion

We have shown the powerful effectiveness of OCR and the necessity of OCR decoupling, which can make a simple network the SOTA. At the same time, we also analyzed the shortcomings of previous work using OCR for retrieval or matching tasks. But there are some problems with our methods as well: **1.** How can we perform text OCR detection without a pre-training model? And existing model is hard to detect text OCR in this kind of text: “**A book is opened to a page that shows pictures of Brisbane in 1895.**”, it is even hard for human beings to distinguish all the OCR, (we cannot make sure whether pictures describe objects or OCR), so how can we solve this arduous task? And this is also the main challenge in the whole field of artificial intelligence. **2.** We just proposed the simplest methods to prove the superiority of OCR. Although the characterization-based method is easy and fast, the error is relatively large. Therefore, methods based on disentangled OCR feature matching will be the next step.

Acknowledgement. This work was partly supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (No. 202200049) and the special project of “Tibet Economic and Social Development and Plateau Scientific Research Co-construction Innovation Foundation” of Wuhan University of Technology&Tibet University (No. lzt2021008).

References

1. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: TextCaps: a dataset for image captioning with reading comprehension. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 742–758. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_44
2. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)
3. Lee, K.-H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 212–228. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_13
4. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: aligning texts with visual concepts. arXiv preprint [arXiv:2111.08276](https://arxiv.org/abs/2111.08276) (2021)
5. Nguyen, M.-D., Nguyen, B.T., Gurrin, C.: A deep local and global scene-graph matching for image-text retrieval. arXiv preprint [arXiv:2106.02400](https://arxiv.org/abs/2106.02400) (2021)
6. Diao, H., et al.: Similarity reasoning and filtration for image-text matching. arXiv preprint [arXiv:2101.01368](https://arxiv.org/abs/2101.01368) (2021)
7. Kant, Y., et al.: Spatially aware multimodal transformers for TextVQA. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 715–732. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_41
8. He, P., et al.: DeBERTa: decoding-enhanced BERT with disentangled attention. arXiv preprint [arXiv:2006.03654](https://arxiv.org/abs/2006.03654) (2020)
9. Devlin, J., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
10. Li, K., et al.: Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
11. Ule, J., et al.: CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**(5648), 1212–1215 (2003)
12. Li, J., et al.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv preprint [arXiv:2201.12086](https://arxiv.org/abs/2201.12086) (2022)
13. Lu, X., Zhao, T., Lee, K.: VisualSparta: an embarrassingly simple approach to large-scale text-to-image search with weighted bag-of-words. arXiv preprint [arXiv:2101.00265](https://arxiv.org/abs/2101.00265) (2021)
14. Messina, N., et al.: Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **17**(4), 1–23 (2021)
15. Mafla, A., et al.: StacMR: scene-text aware cross-modal retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2021)
16. Cheng, M., et al.: ViSTA: vision and scene text aggregation for cross-modal retrieval. arXiv preprint [arXiv:2203.16778](https://arxiv.org/abs/2203.16778) (2022)

17. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *ACL* **2**, 67–78 (2014)
18. 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
19. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: Coco-text: dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140* (2016)
20. Wang, P., et al.: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052* (2022)
21. Biten, A.F., et al.: Is an image worth five sentences? A new look into semantics for image-text matching. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022)
22. Liao, M., Pang, G., Huang, J., Hassner, T., Bai, X.: Mask TextSpotter v3: segmentation proposal network for robust scene text spotting. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *ECCV 2020*. LNCS, vol. 12356, pp. 706–722. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_41
23. Wang, H., et al.: Scene text retrieval via joint text detection and similarity learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021)
24. Ren, S., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems 28* (2015)
25. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: *Advances in Neural Information Processing Systems 28* (2015)
26. Zhu, Q., et al.: Simple is not easy: a simple strong baseline for TextVQA and TextCaps. *arXiv preprint arXiv:2012.05153* 2 (2020)
27. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018)
28. Huang, Z., et al.: Pixel-BERT: aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849* (2020)
29. Miech, A., et al.: Thinking fast and slow: efficient text-to-visual retrieval with transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021)