Efficient Token-Guided Image-Text Retrieval with Consistent Multimodal Contrastive Training

Abstract

Image-text retrieval is a central problem for understanding the semantic relationship between vision and language, and serves as the basis for various visual and language tasks. Most previous works either simply learn coarse-grained representations of the overall image and text, or elaborately establish the correspondence between image regions or pixels and text words. However, the close relations between coarse and fine-grained representations for each modality are important for image-text retrieval but almost neglected. As a result, such previous works inevitably suffer from low retrieval accuracy or heavy computational cost. In this work, we address image-text retrieval from a novel perspective by combining coarse- and fine-grained representation learning into a unified framework. This framework is consistent with human cognition, as humans simultaneously pay attention to the entire sample and regional elements to understand the semantic content. To this end, a Token-Guided Dual Transformer (TGDT) architecture which consists of two homogeneous branches for image and text modalities, respectively, is proposed for image-text retrieval. The TGDT incorporates both coarse- and fine-grained retrievals into a unified framework and beneficially leverages the advantages of both retrieval approaches. A novel training objective called Consistent Multimodal Contrastive (CMC) loss is proposed accordingly to ensure the intra- and inter-modal semantic consistencies between images and texts in the common embedding space. Equipped with a two-stage inference method based on the mixed global and local cross-modal similarity, the proposed method achieves state-of-the-art retrieval performances with extremely low inference time when compared with representative recent approaches.

Keywords: Image-Text Retrieval, multimodal Transformer, Multimodal Contrastive Training

1 Introduction

Image-text retrieval seeks to establish connections between images and texts by leveraging content-based semantic similarities. This encompasses two interrelated tasks: text-to-image retrieval and image-to-text retrieval. In the former, the objective is to identify the image that most closely aligns with the provided text from a set of image candidates. Meanwhile, the latter task involves locating the sentence within the text candidate set that offers the most accurate description of the given image.

The reason why I chose to reproduce a topic on the direction of image- text retrieval is that: It is highly relevant to various computer vision tasks and machine learning approaches, such as image captioning [1], text-to-image synthesis [2], activity understanding [3, 4], multimodal machine translation [5], scene graph generation [6] and zero-shot learning [7, 8]. Notably, this task has garnered significant and sustained attention in both academic literature and industry. However, the persistent challenge lies in the semantic gap between image content and linguistic descriptions, posing a substantial obstacle to the development of practical retrieval systems. Currently, the primary directions in image-text retrieval encompass coarse-grained retrieval, fine-grained retrieval, and vision-language pre-training.

This paper nicely combines the respective advantages of coarse-grained and fine-grained retrieval. Besides, effective cross-modal retrieval approaches with high efficiency are imperative for deployment in realistic industrial scenarios, which the paper also takes into account. To summarize, this paper mainly includes the following four work points:

- It proposes a token-guided dual transformer architecture for image-text retrieval which beneficially leverages the advantages of both coarse- and fine-grained retrieval approaches.
- It introduces a consistent multimodal contrastive loss which could guarantee the separation consistency of distances of both the intra- and inter-modal unpaired samples in the common latent space.
- It presents an effective and efficient inference strategy by sequentially applying global retrieval and local re-ranking in a two-stage manner.
- The methods achieve state-of-the-art performances on several important benchmarks with both high retrieval accuracy and efficiency.

2 Related works

Past endeavors primarily address cross-modal image-text retrieval from three perspectives: 1) The coarse-grained retrieval method directly computes the overall similarity between the input image and full text by mapping the two disparate modalities into a shared embedding space; 2) The fine-grained matching method autonomously aligns image proposals and text fragments by exploring their nuanced cross-modal correspondences; 3) The visual language pre-training (VLP) method supplements the training of VLP models with external data or knowledge sources to enhance the learning of more effective representations.

2.1 Coarse-grained Retrieval Methods

As deep learning advances, the prevalence of end-to-end image-text retrieval continues to grow. Wang et al. [9] used two independent multi-layer perceptrons to process images and texts, and adopted structural features for target optimization. Zheng et al. [10] studied the architecture of two independent CNN networks which process images and text, and used instance loss for target optimization. Faghri et al. [11] proposed a training loss based on hard negative samples mining and triplet sampling. These approaches independently handle global image and text information through two distinct network branches, offering advantages such as rapid reasoning and straightforward pre-calculation. Nevertheless, they fall short in capturing the intricate interactions between object instances and language tokens. The retrieval performance is somewhat constrained, particularly when dealing with complex images and lengthy sentences.

2.2 Fine-grained Retrieval Methods

Fine-grained retrieval usually focuses on the local information of the image and obtains the visual features of the individual by means of object detection. Karpathy et al. [12] extracted features for each image region and text word and aligned them in the common embedded space. Niu et al. [13] emphasized the representation of text, and utilized semantic trees and Recurrent Neural Networks (RNN) to extract text phrase features. Lee et al. [14] introduced multi-layer cross attention between image regions and text words to learn better alignment features. These methods primarily concentrate on aligning fine-grained components of both images and text. Some incorporate cross-attention mechanisms, demanding extensive cross computations across different modalities. While these methods exhibit commendable performance, their inference speed is considerably slow, rendering them impractical for real-world applications.

2.3 Vision-Language Pre-training Methods

Inspired by developments in natural language processing, there is a growing interest in the Vision-Language pretraining model, incorporating external knowledge sources. CLIP [15] and ALIGN [16] use a large number (0.4 and 1.8 Billion) of noise text-image pairs to train the dual encoder to generate representative features for each modality. ViLBERT [17], Oscar [18], VinVL [19] and other methods use large amounts of data on fine-grained images and text elements to train transformer networks. These approaches necessitate extensive external data and well-designed self-supervised tasks to cultivate superior pre-training models, thereby enhancing the model's efficacy across various specific tasks.

3 Method

3.1 Overview

The framework of Token-Guided Dual Transformer is shown in Fig. 1.

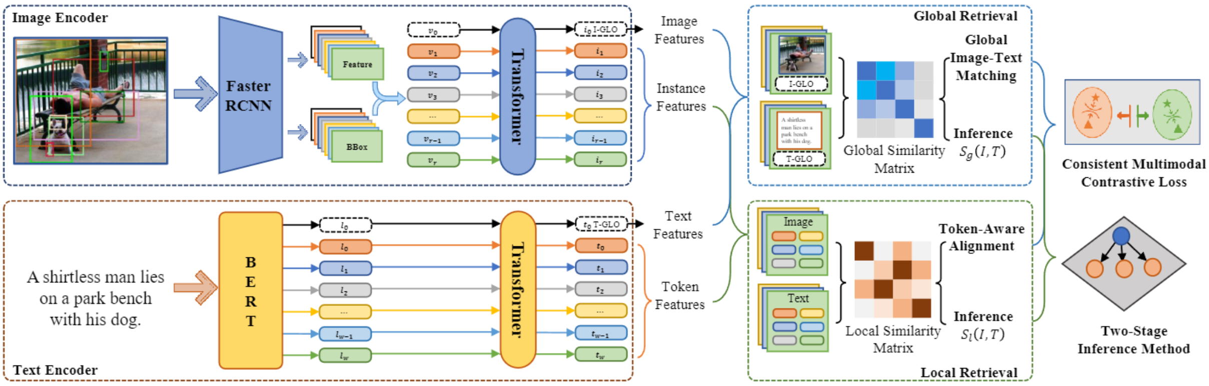


Figure 1. Framework of Token-Guided Dual Transformer (TGDT) architecture.

The paper first describe the transformer-based cross-modal representation learning. Then, it describes both global and local retrievals for image-text retrieval. Finally, the Consistent Multimodal Contrastive (CMC) training loss is introduced, followed by an efficient inference method.

3.2 Cross-Modal Representation Learning

Two transformer encoders are designed to process the image and text modalities, respectively, and collaboratively learn the common feature space.

3.2.1 Image Encoder

With the development of deep learning in computer vision, convolutional neural networks have become the basis of many visual tasks to extract visual information from images. Consistent with [1], and to obtain more descriptive information about visual contents of image regions, it adopts the pre-trained Faster R-CNN [20] as the detector to generate local visual features. For simplicity, it defines the image representation as $V = [v_0, v_1, ..., v_r]$

The transformer encoder learns representations input tokens, and refers to both global and local information at the same time. The paper use a transformer architecture to attend features of both image regions and the whole image. This architecture consists of four identical layers of standard transformer encoder, where each layer is composed of a multi-head self-attention mechanism and a fully connected feed-forward operator. Let $ITR()$ denote the transformer-based image encoder. Each element in V is used as a token input to the transformer head, and the output is learned image representations:$I = ITR(V) = [i_0, i_1, ..., i_r]$

3.2.2 Text Encoder

For text representation, the development of natural language processing has given many excellent representation models. Text can be represented at the sentence or word levels. The paper employs the widely used pre-training model BERT [21] to extract two levels of text semantic information. It defines the text representation as $L = [l_0, l_1, ..., l_r]$

Similarly, the transformer-based text encoder, which is denoted as $TTR()$, also has four identical standard transformer layers. Each element in L is regarded as a token input to this transformer head, and the output are text representation features:$T = TTR(L) = [t_0, t_1, ..., i_w]$

3.3 Joint Global and Local Retrievals

Previous works either use coarse-grained global retrieval or fine-grained local retrieval for image-text retrieval. The proposed TGDT suitably unifies both global and local retrievals under a single framework.

3.3.1 Global Retrieval

The global retrieval only uses global features of the two modalities for cross-modal retrieval. Let $X = [x_0, x_1, ..., x_{nx}]$, $Y = [y_0, y_1, ..., y_{ny}]$ be the learned representations of two samples with different modalities. For example, X is the output features of the image encoder of one sample, and Y is the output features of the text encoder of another sample. Particularly, x_0 and y_0 represent the global features of the two samples in the common feature space. The cosine similarity is used to measure the similarity between two samples. For any two samples, which are represented by X and Y, respectively, the global cross-modal similarity is defined as:

$$S_g(X, Y) = \frac{x_0^T \times y_0}{\|x_0\| \times \|y_0\|}, \tag{1}$$

3.3.2 Local Retrieval

The local retrieval fully utilize local features of the two modalities for retrieval. It calculates the similarity by aligning local elements between the two samples. Let $X = [x_1, ..., x_{nx}]$ and $Y = [y_1, ..., y_{ny}]$ be the local features in the common feature space of image and text modalities, respectively. It aligns samples by calculating the cosine similarity between each element:

$$M_{ij}(X, Y) = \frac{x_i^T \times y_j}{\|x_i\| \times \|y_j\|} \tag{2}$$

Assume that X is the local image features of one sample and Y is the local text features of another sample, the local similarity is defined as:

$$S_l(X, Y) = \frac{1}{n_y} \sum_{j \in [1, n_y]} \max_{i \in [1, n_x]} M_{ij}(X, Y) \tag{3}$$

The advantage of local similarity is that more refined features achieve better retrieval performance. But the crossover between image and text increases the amount of calculation during retrieval, especially for the cross-attention based and highly entangled approaches which require inefficient finegrained alignment between words and image regions.

3.4 Consistent Multimodal Contrastive Loss

3.4.1 Multimodal Contrastive Loss

For image-text retrieval, the triplet ranking loss has been widely used in previous works [22, 23]. This loss is formally defined as:

$$\begin{aligned} L_{\mathrm{r}} &= \max\left(0, \delta - S(I, T) + S\left(I, T_{l^-}\right)\right) \\ &+ \max\left(0, \delta - S(I, T) + S\left(I_{v^-}, T\right)\right) \end{aligned} \tag{4}$$

3.4.2 Consistent Multimodal Contrastive Loss

The basic intuition behind the proposed Consistent Multimodal Contrastive (CMC) loss is to consider the matched image-text pair as a compact sample and try to ensure the consistency of the global distance between different pairs. Given the three image-text pairs $(I, T), (I_{v^-}, T_{v^-}) and (I_{l^-}, T_{l^-})$ mentioned above, the paper first define the contrastive loss within the same modality, whose definition is as follows:

$$\begin{aligned} L_{\mathrm{a}} &= \max\left(0, \left|S\left(I, I_{l^-}\right) - S\left(T, T_{l^-}\right)\right| - \sigma\right) \\ &+ \max\left(0, \left|S\left(I, I_{v^-}\right) - S\left(T, T_{v^-}\right)\right| - \sigma\right) \end{aligned} \tag{5}$$

The proposed CMC loss can be directly obtained by adding the inter-modal loss and the intra-modal loss:

$$L_{\mathrm{cmc}} = L_{\mathrm{r}} + L_{\mathrm{a}}. \tag{6}$$

4 Implementation details

4.1 Comparing with the released source codes

In the reproduction process, I refer to the source code provided by TGDT for reproduction:github.com/LCFractal/TGDT

On the basis of TGDT, I also consider the problem of bounding box accuracy in fine-grained retrieval, which may affect the overall retrieval effect.Therefore, I refer to the Bounding Box Prediction in X-VLM [24] and introduce it into the training process.

4.2 Experimental environment setup

For the image input, the paper use the object proposals provided by [1], which selects the top 36 region proposals with the highest confidence scores, and describes each object proposal with a 2048-dimensional bottom-up feature vector. For the text input, it use the BERT model which is pre-trained on the mask language task of English sentences to obtain the 768-dimensional embedding features of the text. For transformer encoders of both modalities, the embedding size of self-attention layers is 1024, and the output feature dimension is 2048. The default margin or slack hyperparameters δ = 0.2 and σ = 0.3 in the proposed CMC loss function are 0.2 and 0.3, respectively. During training, it use Adam as the optimizer. The initial learning rate, number of training epochs, and batch size are 1e 6, 30 and 40, respectively.

4.3 Main contributions

The innovation I did was to introduce Bounding Box Prediction into the training process, and I followed the practice in X-VLM [24]. Letting the model predict the bounding box b^j of visual concept V^j given the image representation and the text representation, where $b^j = (cx, cy, w, h)$. By locating different visual concepts in the same image, it is expected that the model better learns finegrained vision language alignments. The bounding box is predicted by:

$$\hat{\boldsymbol{b}}^j\left(I, T^j\right)=\operatorname{Sigmoid}\left(\operatorname{MLP}\left(x_{\text{cls}}^j\right)\right) \tag{7}$$

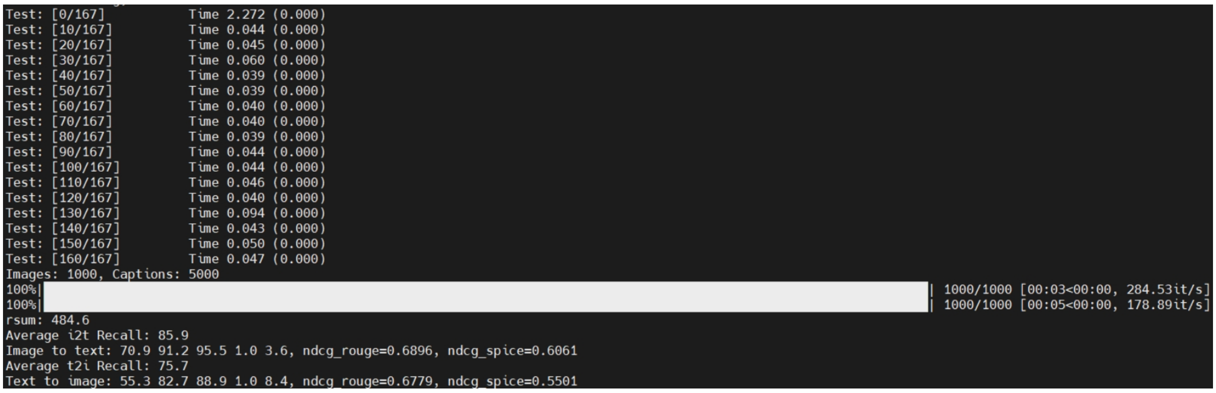
For bounding box prediction, L_1 is the most commonly-used loss. However, it has different scales for small and large boxes, even if their relative errors are similar. To mitigate this issue, we use a linear combination of the L_1 loss and the generalized Intersection over Union (IoU) loss [25], which is scale-invariant. The overall loss is defined as:

$$\mathcal{L}_{\text{bbox}}=\mathbb{E}_{(V^j,T^j)\sim I;I\sim D}\left[\mathcal{L}_{\text{iou}}\left(\boldsymbol{b}_j,\hat{\boldsymbol{b}}_j\right)+\left\|\boldsymbol{b}_j-\hat{\boldsymbol{b}}_j\right\|_1\right] \tag{8}$$

5 Results and analysis

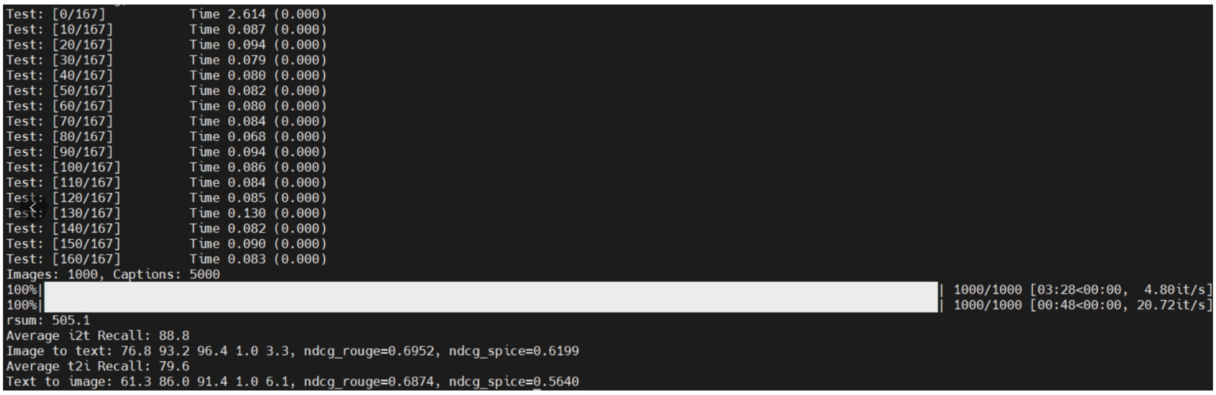
First of all, according to the Settings in the TGDT source code, I successfully reproduce the image text retrieval results on two datasets MSCOCO and Flickr30k in the paper.

Flickr30k:



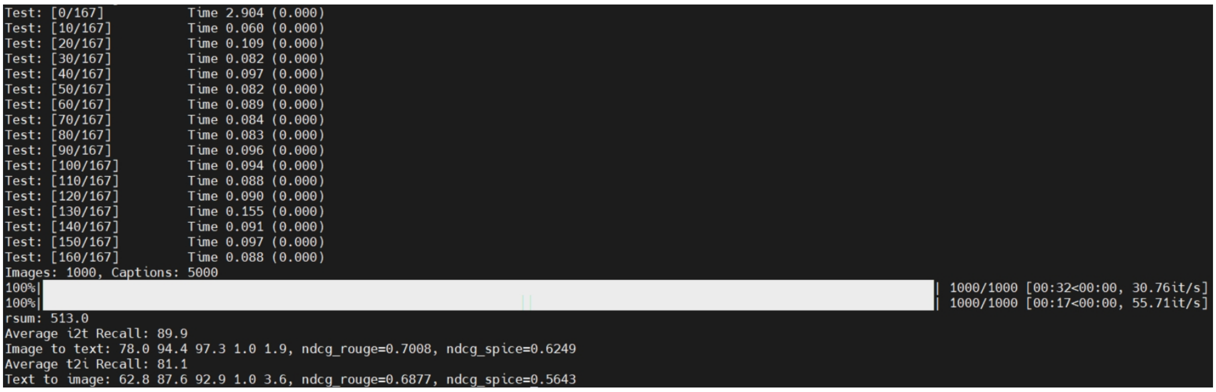
```
Test: [0/167]           Time 2.272 (0.000)
Test: [10/167]          Time 0.044 (0.000)
Test: [20/167]          Time 0.045 (0.000)
Test: [30/167]          Time 0.060 (0.000)
Test: [40/167]          Time 0.039 (0.000)
Test: [50/167]          Time 0.039 (0.000)
Test: [60/167]          Time 0.040 (0.000)
Test: [70/167]          Time 0.040 (0.000)
Test: [80/167]          Time 0.039 (0.000)
Test: [90/167]          Time 0.044 (0.000)
Test: [100/167]         Time 0.044 (0.000)
Test: [110/167]         Time 0.046 (0.000)
Test: [120/167]         Time 0.040 (0.000)
Test: [130/167]         Time 0.094 (0.000)
Test: [140/167]         Time 0.043 (0.000)
Test: [150/167]         Time 0.050 (0.000)
Test: [160/167]         Time 0.047 (0.000)
Images: 1000, Captions: 5000
100%|                                                                                | 1000/1000 [00:03<00:00, 284.53it/s]
100%|                                                                                | 1000/1000 [00:05<00:00, 178.89it/s]
rsum: 484.6
Average i2t Recall: 85.9
Image to text: 70.9 91.2 95.5 1.0 3.6, ndcg_rouge=0.6896, ndcg_spice=0.6061
Average t2i Recall: 75.7
Text to image: 55.3 82.7 88.9 1.0 8.4, ndcg_rouge=0.6779, ndcg_spice=0.5501
```

Figure 2. Experimental results of Flickr30k-global



```
Test: [0/167]           Time 2.614 (0.000)
Test: [10/167]          Time 0.087 (0.000)
Test: [20/167]          Time 0.094 (0.000)
Test: [30/167]          Time 0.079 (0.000)
Test: [40/167]          Time 0.080 (0.000)
Test: [50/167]          Time 0.082 (0.000)
Test: [60/167]          Time 0.080 (0.000)
Test: [70/167]          Time 0.084 (0.000)
Test: [80/167]          Time 0.068 (0.000)
Test: [90/167]          Time 0.094 (0.000)
Test: [100/167]         Time 0.086 (0.000)
Test: [110/167]         Time 0.084 (0.000)
Test: [120/167]         Time 0.085 (0.000)
Test: [130/167]         Time 0.130 (0.000)
Test: [140/167]         Time 0.082 (0.000)
Test: [150/167]         Time 0.090 (0.000)
Test: [160/167]         Time 0.083 (0.000)
Images: 1000, Captions: 5000
100%|                                                                                | 1000/1000 [03:28<00:00,  4.80it/s]
100%|                                                                                | 1000/1000 [00:48<00:00, 20.72it/s]
rsum: 505.1
Average i2t Recall: 88.8
Image to text: 76.8 93.2 96.4 1.0 3.3, ndcg_rouge=0.6952, ndcg_spice=0.6199
Average t2i Recall: 79.6
Text to image: 61.3 86.0 91.4 1.0 6.1, ndcg_rouge=0.6874, ndcg_spice=0.5640
```

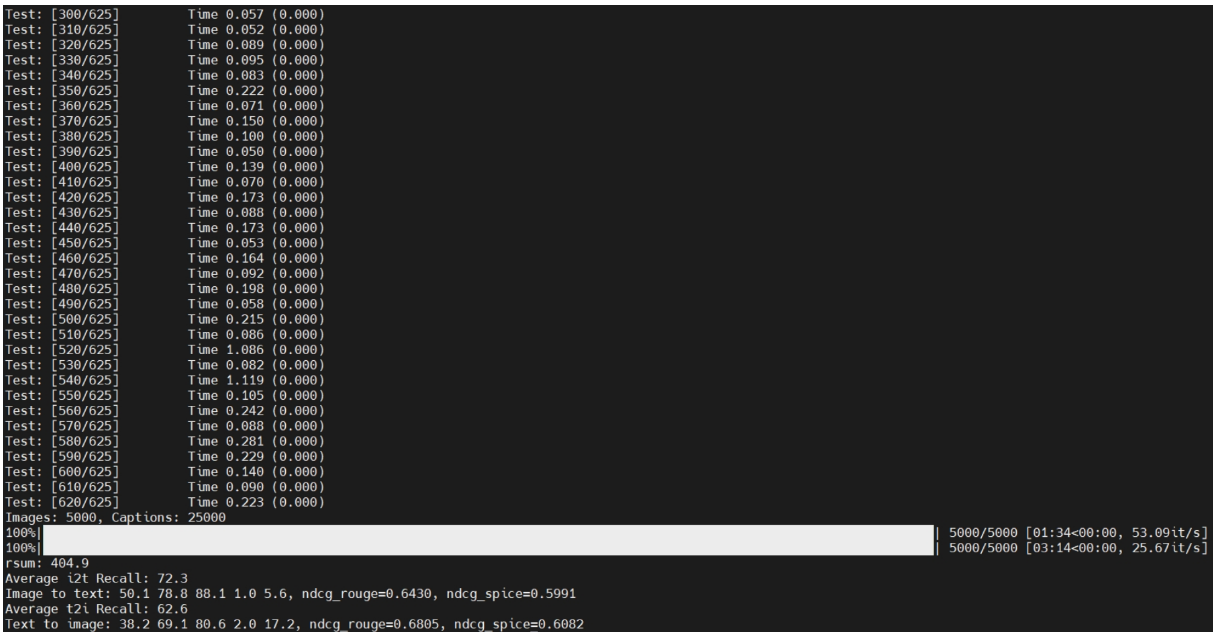
Figure 3. Experimental results of Flickr30k-local



```
Test: [0/167]           Time 2.904 (0.000)
Test: [10/167]          Time 0.060 (0.000)
Test: [20/167]          Time 0.109 (0.000)
Test: [30/167]          Time 0.082 (0.000)
Test: [40/167]          Time 0.097 (0.000)
Test: [50/167]          Time 0.082 (0.000)
Test: [60/167]          Time 0.089 (0.000)
Test: [70/167]          Time 0.084 (0.000)
Test: [80/167]          Time 0.083 (0.000)
Test: [90/167]          Time 0.096 (0.000)
Test: [100/167]         Time 0.094 (0.000)
Test: [110/167]         Time 0.088 (0.000)
Test: [120/167]         Time 0.090 (0.000)
Test: [130/167]         Time 0.155 (0.000)
Test: [140/167]         Time 0.091 (0.000)
Test: [150/167]         Time 0.097 (0.000)
Test: [160/167]         Time 0.088 (0.000)
Images: 1000, Captions: 5000
100%|                                                                                | 1000/1000 [00:32<00:00, 30.76it/s]
100%|                                                                                | 1000/1000 [00:17<00:00, 55.71it/s]
rsum: 513.0
Average i2t Recall: 89.9
Image to text: 78.0 94.4 97.3 1.0 1.9, ndcg_rouge=0.7008, ndcg_spice=0.6249
Average t2i Recall: 81.1
Text to image: 62.8 87.6 92.9 1.0 3.6, ndcg_rouge=0.6877, ndcg_spice=0.5643
```

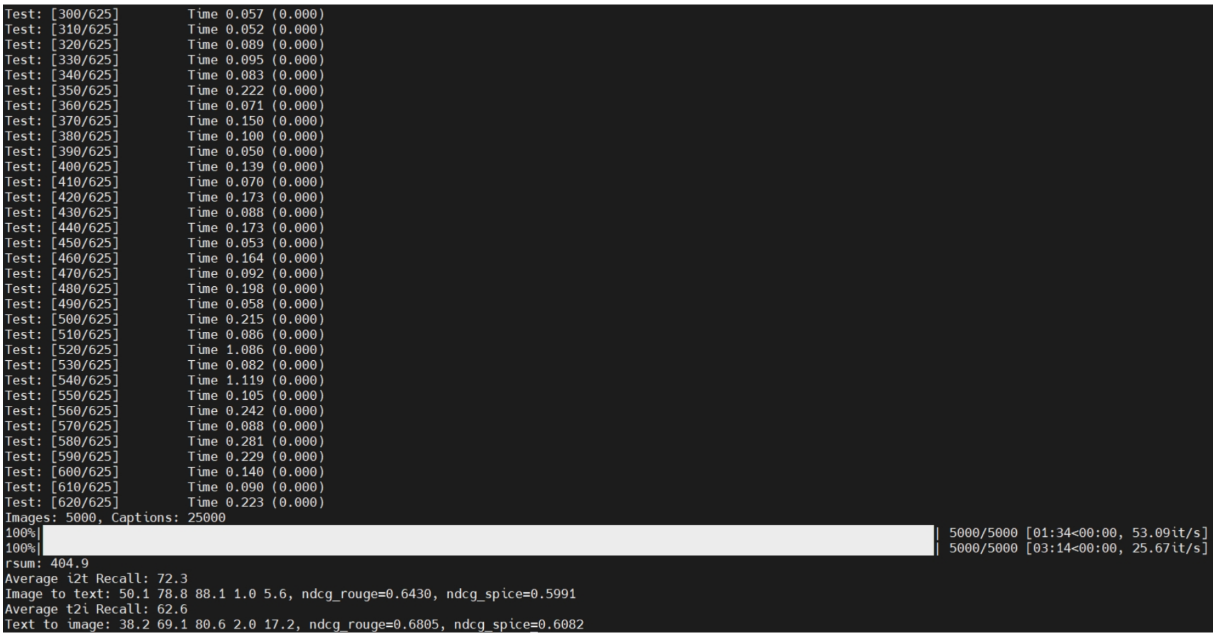
Figure 4. Experimental results of Flickr30k-global-local

MSCOCO:



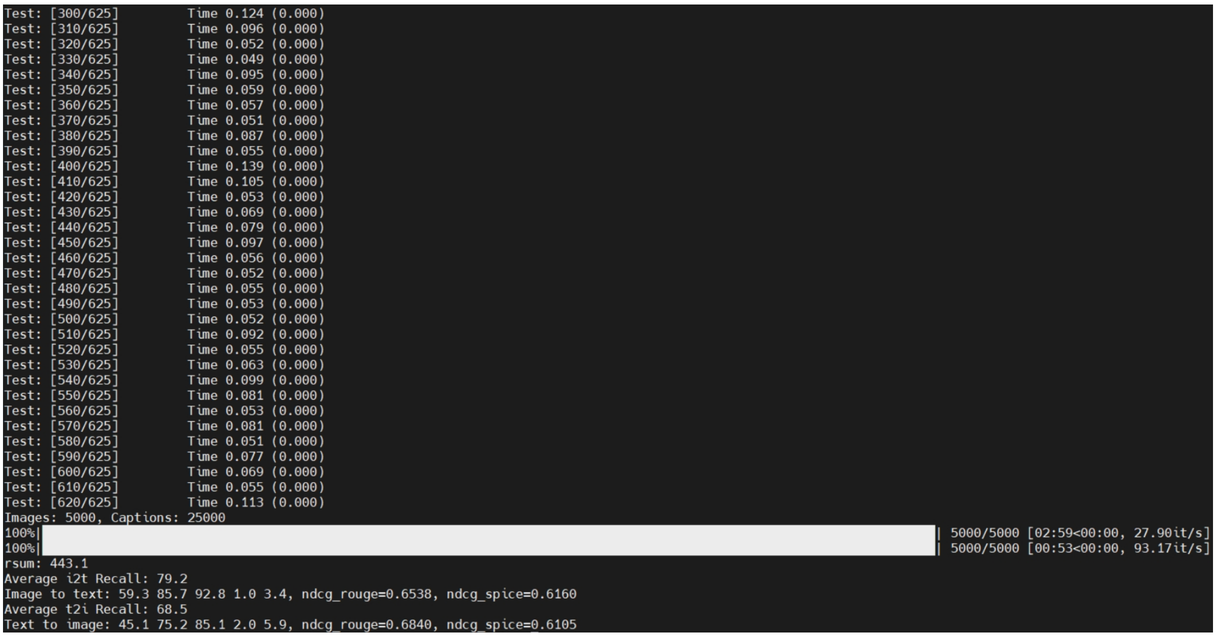
```
Test: [300/625]        Time 0.057 (0.000)
Test: [310/625]        Time 0.052 (0.000)
Test: [320/625]        Time 0.089 (0.000)
Test: [330/625]        Time 0.095 (0.000)
Test: [340/625]        Time 0.083 (0.000)
Test: [350/625]        Time 0.222 (0.000)
Test: [360/625]        Time 0.071 (0.000)
Test: [370/625]        Time 0.150 (0.000)
Test: [380/625]        Time 0.100 (0.000)
Test: [390/625]        Time 0.050 (0.000)
Test: [400/625]        Time 0.139 (0.000)
Test: [410/625]        Time 0.070 (0.000)
Test: [420/625]        Time 0.173 (0.000)
Test: [430/625]        Time 0.088 (0.000)
Test: [440/625]        Time 0.173 (0.000)
Test: [450/625]        Time 0.053 (0.000)
Test: [460/625]        Time 0.164 (0.000)
Test: [470/625]        Time 0.092 (0.000)
Test: [480/625]        Time 0.198 (0.000)
Test: [490/625]        Time 0.058 (0.000)
Test: [500/625]        Time 0.215 (0.000)
Test: [510/625]        Time 0.086 (0.000)
Test: [520/625]        Time 1.086 (0.000)
Test: [530/625]        Time 0.082 (0.000)
Test: [540/625]        Time 1.119 (0.000)
Test: [550/625]        Time 0.105 (0.000)
Test: [560/625]        Time 0.242 (0.000)
Test: [570/625]        Time 0.088 (0.000)
Test: [580/625]        Time 0.281 (0.000)
Test: [590/625]        Time 0.229 (0.000)
Test: [600/625]        Time 0.140 (0.000)
Test: [610/625]        Time 0.090 (0.000)
Test: [620/625]        Time 0.223 (0.000)
Images: 5000, Captions: 25000
100%|                                                  | 5000/5000 [01:34<00:00, 53.09it/s]
100%|                                                  | 5000/5000 [03:14<00:00, 25.67it/s]
rsum: 404.9
Average i2t Recall: 72.3
Image to text: 50.1 78.8 88.1 1.0 5.6, ndcg_rouge=0.6430, ndcg_spice=0.5991
Average t2i Recall: 62.6
Text to image: 38.2 69.1 80.6 2.0 17.2, ndcg_rouge=0.6805, ndcg_spice=0.6082
```

Figure 5. Experimental results of MSCOCO-global



```
Test: [300/625]        Time 0.057 (0.000)
Test: [310/625]        Time 0.052 (0.000)
Test: [320/625]        Time 0.089 (0.000)
Test: [330/625]        Time 0.095 (0.000)
Test: [340/625]        Time 0.083 (0.000)
Test: [350/625]        Time 0.222 (0.000)
Test: [360/625]        Time 0.071 (0.000)
Test: [370/625]        Time 0.150 (0.000)
Test: [380/625]        Time 0.100 (0.000)
Test: [390/625]        Time 0.050 (0.000)
Test: [400/625]        Time 0.139 (0.000)
Test: [410/625]        Time 0.070 (0.000)
Test: [420/625]        Time 0.173 (0.000)
Test: [430/625]        Time 0.088 (0.000)
Test: [440/625]        Time 0.173 (0.000)
Test: [450/625]        Time 0.053 (0.000)
Test: [460/625]        Time 0.164 (0.000)
Test: [470/625]        Time 0.092 (0.000)
Test: [480/625]        Time 0.198 (0.000)
Test: [490/625]        Time 0.058 (0.000)
Test: [500/625]        Time 0.215 (0.000)
Test: [510/625]        Time 0.086 (0.000)
Test: [520/625]        Time 1.086 (0.000)
Test: [530/625]        Time 0.082 (0.000)
Test: [540/625]        Time 1.119 (0.000)
Test: [550/625]        Time 0.105 (0.000)
Test: [560/625]        Time 0.242 (0.000)
Test: [570/625]        Time 0.088 (0.000)
Test: [580/625]        Time 0.281 (0.000)
Test: [590/625]        Time 0.229 (0.000)
Test: [600/625]        Time 0.140 (0.000)
Test: [610/625]        Time 0.090 (0.000)
Test: [620/625]        Time 0.223 (0.000)
Images: 5000, Captions: 25000
100%|                                                  | 5000/5000 [01:34<00:00, 53.09it/s]
100%|                                                  | 5000/5000 [03:14<00:00, 25.67it/s]
rsum: 404.9
Average i2t Recall: 72.3
Image to text: 50.1 78.8 88.1 1.0 5.6, ndcg_rouge=0.6430, ndcg_spice=0.5991
Average t2i Recall: 62.6
Text to image: 38.2 69.1 80.6 2.0 17.2, ndcg_rouge=0.6805, ndcg_spice=0.6082
```

Figure 6. Experimental results of MSCOCO-local



```
Test: [300/625]         Time 0.124 (0.000)
Test: [310/625]         Time 0.096 (0.000)
Test: [320/625]         Time 0.052 (0.000)
Test: [330/625]         Time 0.049 (0.000)
Test: [340/625]         Time 0.095 (0.000)
Test: [350/625]         Time 0.059 (0.000)
Test: [360/625]         Time 0.057 (0.000)
Test: [370/625]         Time 0.051 (0.000)
Test: [380/625]         Time 0.087 (0.000)
Test: [390/625]         Time 0.055 (0.000)
Test: [400/625]         Time 0.139 (0.000)
Test: [410/625]         Time 0.105 (0.000)
Test: [420/625]         Time 0.053 (0.000)
Test: [430/625]         Time 0.069 (0.000)
Test: [440/625]         Time 0.079 (0.000)
Test: [450/625]         Time 0.097 (0.000)
Test: [460/625]         Time 0.056 (0.000)
Test: [470/625]         Time 0.052 (0.000)
Test: [480/625]         Time 0.055 (0.000)
Test: [490/625]         Time 0.053 (0.000)
Test: [500/625]         Time 0.052 (0.000)
Test: [510/625]         Time 0.092 (0.000)
Test: [520/625]         Time 0.055 (0.000)
Test: [530/625]         Time 0.063 (0.000)
Test: [540/625]         Time 0.099 (0.000)
Test: [550/625]         Time 0.081 (0.000)
Test: [560/625]         Time 0.053 (0.000)
Test: [570/625]         Time 0.081 (0.000)
Test: [580/625]         Time 0.051 (0.000)
Test: [590/625]         Time 0.077 (0.000)
Test: [600/625]         Time 0.069 (0.000)
Test: [610/625]         Time 0.055 (0.000)
Test: [620/625]         Time 0.113 (0.000)
Images: 5000, Captions: 25000
100%|████████████████████████████████████████| 5000/5000 [02:59<00:00, 27.90it/s]
100%|████████████████████████████████████████| 5000/5000 [00:53<00:00, 93.17it/s]
rsum: 443.1
Average i2t Recall: 79.2
Image to text: 59.3 85.7 92.8 1.0 3.4, ndcg_rouge=0.6538, ndcg_spice=0.6160
Average t2i Recall: 68.5
Text to image: 45.1 75.2 85.1 2.0 5.9, ndcg_rouge=0.6840, ndcg_spice=0.6105
```

Figure 7. Experimental results of MSCOCO-global-local

The reproducible results, mirroring the outcomes detailed in the paper, affirm the model's robust performance across varied datasets. Notably, we can clearly see from the results of the model on two different datasets that the accuracy of global-local is higher considering both image retrieval of text and text retrieval of image. The demonstrated complementarity of global and local representations highlights the nuanced interplay between these aspects in achieving superior retrieval accuracy. This suggests that a holistic approach, considering both global and local features, contributes significantly to the model's proficiency in image-text retrieval tasks.

6 Conclusion and future work

Having thoroughly reviewed both the paper TGDT and its source code, I successfully reproduce the work. Additionally, I put forth the notion of incorporating Bounding Box Prediction, a novel approach aimed at enhancing the overall performance of image-text retrieval models. While this introduces a promising avenue for improvement, there are existing challenges, notably the prolonged training time and the need for further refinement in accuracy rates. Looking ahead, there is ample room for future research to delve into optimizing the efficiency and accuracy of the model. Addressing these aspects will contribute to advancing the state-of-the-art in image-text retrieval, paving the way for more effective and streamlined applications in diverse domains.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086, 2018.

[2] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[3] Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Zhihui Li, Lina Yao, and Alex Hauptmann. Tnzstad: Transferable network for zero-shot temporal activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3848–3861, 2022.

[4] Hongsong Wang and Liang Wang. Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. *IEEE Transactions on Image Processing*, 27(9):4382–4394, 2018.

[5] Mingjie Li, Po-Yao Huang, Xiaojun Chang, Junjie Hu, Yi Yang, and Alex Hauptmann. Video pivoting unsupervised multi-modal machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3918–3932, 2022.

[6] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1–26, 2021.

[7] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[8] Caixia Yan, Xiaojun Chang, Zhihui Li, Weili Guan, Zongyuan Ge, Lei Zhu, and Qinghua Zheng. Zeronas: Differentiable generative adversarial networks search for zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):9733–9740, 2021.

[9] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[10] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.

[11] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[12] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[13] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 1881–1889, 2017.

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

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[16] 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 *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[17] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[18] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[19] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[22] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018.

[23] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2840–2848, 2017.

[24] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

[25] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.