VisualSparta: Sparse Transformer Fragment-level Matching for Large-scale Text-to-Image Search

Xiaopeng Lu
Language Technologies Institute
Carnegie Mellon University
xiaopen2@andrew.cmu.edu

Tiancheng Zhao, Kyusong Lee
SOCO Inc
Pittsburgh, USA
{tianchez,kyusongl}@soco.ai

Abstract

Text-to-image retrieval is an essential task in multi-modal information retrieval, i.e. retrieving relevant images from a large and unlabelled image dataset given textual queries. In this paper, we propose VisualSparta, a novel text-to-image retrieval model that shows substantial improvement over existing models on both accuracy and efficiency. We show that VisualSparta is capable of outperforming all previous scalable methods in MSCOCO and Flickr30K. It also shows substantial retrieving speed advantages, i.e. for an index with 1 million images, VisualSparta gets over 391x speed up compared to standard vector search. Experiments show that this speed advantage even gets bigger for larger datasets because VisualSparta can be efficiently implemented as an inverted index. To the best of our knowledge, VisualSparta is the first transformer-based text-to-image retrieval model that can achieve real-time searching for very large dataset, with significant accuracy improvement compared to previous state-of-the-art methods.

1 Introduction

Text-to-image retrieval is the task of retrieving the most relevant images given a text query. Given a text description, the model needs to output a list of relevant images from the indexed corpus. In this paper, we argue that current text-to-image retrieval models face two main challenges: accuracy challenge and latency challenge. We further propose VisualSparta, an accurate and efficient retrieval model that significantly outperforms other existing models on both aspects.

Achieving high accuracy is challenging in text-to-image retrieval task. In order to find the most relevant images given text query, the model needs to not only have good representations for both text and image modalities, but also understand the fine-grained relationships between two representations. For example, given a query "A girl in black jacket drinking milk and eating pizza.", the model needs to first have good representations for the important concepts such as "girl", "black jacket", "milk". For each image, the model has to capture the large amounts of fine-grained information lying in it. Then, the model needs to learn to pick the most relevant images given the text query. Much work has been done on text-to-image retrieval task. Some methods rely on a dual-encoder architecture, which uses two encoder to separately encode text query and image answer, then optimize the encoder weights for both sides (Faghri *et al.*, 2017; Lee *et al.*, 2018; Wang *et al.*, 2019a). Other recent methods leverage an transformer-based architecture (Devlin *et al.*, 2018; Yang *et al.*, 2019). In this case, each pair of text and image is encoded by concatenating and passing into one single network, instead of encoded by two separate encoders (Lu *et al.*, 2020; Li *et al.*, 2020). This transformer-based method borrows the knowledge from large pretrained models and shows better results compared to dual-encoder method.

Another challenge that previous works did not touch upon is the retrieval latency challenge. Retrieval latency is a long-existing challenge in information retrieval (IR) area, as the design of an IR system needs to fit user's real-time information seeking needs (Manning *et al.*, 2008). Although text information retrieval communities start focusing on this problem for a long time and view latency as an important metric, latency problem has not been well studied in text-to-image retrieval problems yet. In this paper, we evaluate and analyze the speed performance of our method, and do detailed speed comparisons with existing methods.

In this work, we propose VisualSparta (Sparse Transformer Fragment-level Matching), a simple yet effective text-to-image retrieval model that per-

forms better than all existing retrieval models on both accuracy and retrieval latency. By modeling fragment-level information for the query and region information for the answer, our model is capable of borrowing information from large pretrained models, while benefiting from the efficiency of sparse retrieval method. To the best of our knowledge, this is the first model that integrates the power of transformer-based models with the real-time searching capabilities, showing that large pretrained models can be used in a way that uses significantly less amount of memory and computing time.

Contributions of this paper can be concluded as the following:

- A novel image-to-text retrieval model that based on fragment-level interaction and shows state-of-the-art results over two benchmark datasets.
- Inverted index are shown to be effective in text-to-image search and produces promising results in terms of accuracy and efficiency.
- Detailed analysis on accuracy-latency comparisons over current text-to-image retrieval models to show the strong performance of our proposed model in large-scale settings

2 Related Work

Large amounts of work have been done on learning a joint representation between texts and images (Karpathy and Fei-Fei, 2015; Huang *et al.*, 2018; Lee *et al.*, 2018; Wehrmann *et al.*, 2019; Li *et al.*, 2020; Lu *et al.*, 2020). In this section, we revisit dual-encoder based retrieval model and transformer-based retrieval model.

2.1 Dual-encoder Matching Network

Most of the work in text-to-image retrieval task choose to use dual-encoder network to encode information from text and image modalities. In Karpathy and Fei-Fei (2015), the author used a Bi-directional Recurrent Neural Network (BRNN) to encode the textual information and used an Region Convolutional Neural Netork (RCNN) to encode the image information, and the final similarity score is computed via the interaction of features from two encoders. Lee *et al.* (2018) proposed stacked cross-attention network, where the text features are passed through two attention layers to learn interactions with the image region. Wang *et al.* (2019a) encoded the location information as yet another feature, and used both deep Faster-RCNN features (Ren *et al.*, 2016) and the fine-grained location features for the Region of Interest (ROI) as image representation. In Wang *et al.* (2020), the author utilized the information from external corpus (Wikipedia) to construct a Graph Neural Network (GNN) to help model the relationships across objects.

2.2 Pretrained Language Models (PLM)

Large pretrained language models (PLM) show great success over multiple tasks in NLP areas in recent years (Devlin *et al.*, 2018; Yang *et al.*, 2019; Dai *et al.*, 2019). After that, research has also been done on cross-modal transformer-based models and proves that the self-attention mechanism also helps jointly capture text-image relationships (Li *et al.*, 2019; Lu *et al.*, 2020; Qi *et al.*, 2020; Li *et al.*, 2020). By first pretraining model under large-scale text-image dataset, these transformer-based models capture rich semantic information from both texts and images. Models are then fine-tuned for the text-to-image retrieval task and show improvements by a large margin. However, the problem of using transformer-based models is that it is prohibitively slow in the retrieval context: the model needs to compute pair-wise similarity scores between all queries and answers, making it almost impossible to use the model in any real-world scenarios. Our proposed method borrows the power of large pretrained models, while reducing the inference time by orders of magnitude.

PLM has shown promising results in Information Retrieval (IR), despite of its prohibitively slow speed due to the complex model structure. IR communities has recently started working on empowering the classical full-text retrieval methods with contextualized information from PLMs (Dai and Callan, 2019; MacAvaney *et al.*, 2020; Zhao *et al.*, 2020). Dai and Callan (2019) proposed DeepCT, a model which learns to generate the query importance score from the contextualized representation of large transformer-based models. Zhao *et al.* (2020) proposed sparse transformer matching model, where the model learns term-level interaction between query and text answers and generates weighted term representations for answers during index time. Our work is motivated by works in this direction, and extends the scope to the cross-modal

understanding and retrieval.

3 VisualSparta Retriever

In this section, we present VisualSparta retriever, a fragment-level transformer-based model for efficient text-image matching. The focus of our proposed model is two-fold:

- Accuracy: fine-grained relationship between query tokens and image regions are learned to enrich the cross-modal understanding.
- Efficiency: learning query and answer embedding independently allows the model to index all the images offline, and the whole VisualSparta model can be adopted to a classical inverted-index search engine for efficient search.

3.1 Model Architecture

3.1.1 Query representation

As the query processing is an online operation during retrieval, efficiency of encoding query needs to be well considered. Previous methods pass the query sentence into an bi-RNN to give token representation given surrounding tokens (Lee *et al.*, 2018; Wang *et al.*, 2019a, 2020).

Instead of encoding the query in a sequential manner, we drop the order information of the query and only use the pretrained word embedding to represent each token. In other words, we do not encode the local contextual information for the query, and purely rely on independent word embedding E_{word} of each token. Let a query be $q = w_1, ..., w_m$:

$$\hat{w}_i = E_{word}(w_i) \tag{1}$$

where w_i is the i-th word of the query. Therefore, a query is represented as $\hat{w} = \{\hat{w}_1, ..., \hat{w}_m\}, \hat{w}_i \in \mathbb{R}^{d_H}$. In this way, each token can be represented independently and agnostic to the local context. This is essential for the efficient indexing and inference, as described next in 3.3.

3.1.2 Visual Representation

Compared with query which needs to be processed at real time, answer processing can be rich and complex, as answer corpus can be indexed offline before the query comes. Therefore, we follow the previous work (Li *et al.*, 2020) and use the contextualized representation for the answer corpus.

Specifically, for an image, we represent it using information from three sources: regional deep features, regional location features, and object label features.

Regional deep features and location features Given an image v, we pass it through Faster-RCNN (Ren *et al.*, 2016) to get n regional deep features v_i and their corresponding location features l_i:

$$v_1, ..., v_n = \text{Faster-RCNN}(v), v_i \in \mathbb{R}^{d_{rcnn}} \tag{2}$$

and the location features are the normalized top left and bottom right right positions of the region proposed from Faster-RCNN, plus the region width and height:

$$l_i = [l_{xmin}, l_{xmax}, l_{ymin}, l_{ymax}, l_{width}, l_{height}] \tag{3}$$

Therefore, we represent one region by the concatenation of two features:

$$E_i = [v_i; l_i] \tag{4}$$

$$E_{image} = [E_1, ..., E_n], E_i \in \mathbb{R}^{d_{rcnn}+d_{loc}} \tag{5}$$

where E_{image} is the representation for a single image.

Object label features Additional to the deep representations from the proposed image region, previous work (Li *et al.*, 2020) shows that the object label information is also useful as an additional representation for the image. We also encode the predicted objects obtained from Faster-RCNN model with pretrained word embeddings:

$$\hat{o_i} = E_{word}(o_i) + E_{pos}(o_i) + E_{seg}(o_i) \tag{6}$$

$$E_{label} = [\hat{o_1}, ..., \hat{o_k}], \hat{o_i} \in \mathbb{R}^{d_H} \tag{7}$$

where o_i represents one object label, and $E_{word}, E_{pos}, E_{seg}$ represent word embedding, position embedding, and segment embeddings respectively, same as the embedding structure in Devlin *et al.* (2018).

Therefore, one image can be represented as linear transformed image features concatenated with label features:

$$a = [(E_{image}W + b); E_{label}] \tag{8}$$

where $W \in \mathbb{R}^{(d_{rcnn}+d_{loc}) \times d_H}$ and $b \in \mathbb{R}^{d_H}$ are the trainable linear combination weights and bias. The concatenated a are then passed into a Transformer

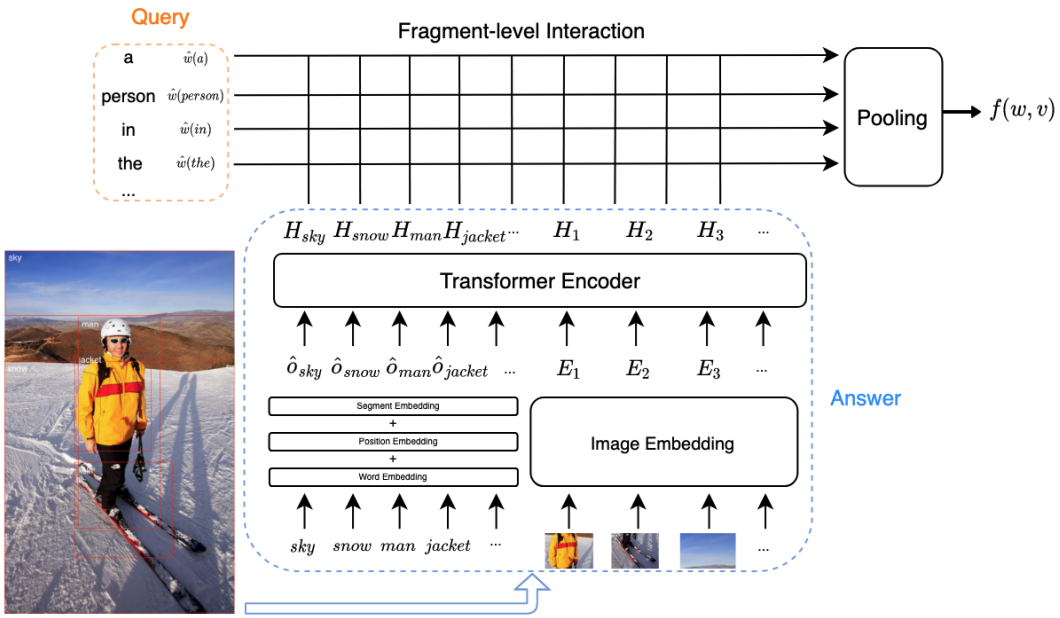

Figure 1: Visual-SPARTA Model. It first computes contextual image region representation and query token representation. Then it computes matching score between every query token and image region that can be stored in an inverted index for efficient searching.

encoder T_{image}, and the final image feature is the hidden output of it:

$$H_{image} = T_{image}(a) \tag{9}$$

where $H_{image} \in \mathbb{R}^{(n+k)\times d_H}$ is the final contextualized representation for one answer.

3.1.3 Scoring Function

Given the visual and query representations, now we are ready to compute the matching score between a query and an image. Different from other dual-encoder based interaction model, we adopt the fine-grained interaction model proposed by (Zhao *et al.*, 2020) to to compute the relevance score by:

$$y_i = \max_{j\in[1,n+k]}(\hat{w}_i^T h_j) \tag{10}$$

$$\phi(y_i) = \text{ReLU}(y_i + b) \tag{11}$$

$$f(q, v) = \sum_{i=0}^{m} \log(\phi(y_i) + 1) \tag{12}$$

where Eq.10 captures the fragment-level interaction between every image element and every query word token; Eq.11 produces sparse embedding outputs via a combination of ReLu and trainable bias and Eq.12 sums up the score and prevents overly large score using log operation.

3.2 Retriever training

Following training method (Zhao *et al.*, 2020), we use cross entropy loss to train VisualSparta. Concretely, we maximize the objective in Eq. 13, which tries to decide between the ground truth image v^+ and irrelevant/random images V^- for each text query q. The parameters to learn include both the query encoder E_{word} and the image transformer encoder T_{image}. Parameters are optimized using Adam (Kingma and Ba, 2014).

$$J = f(q, v^+) - \log \sum_{k\in K^-} e^{f(q,v_k))} \tag{13}$$

In order to achieve efficient training, we use other image samples from the same batch as negative examples for each training data, an effective technique that is widely used in response selection (Zhang *et al.*, 2018; Henderson *et al.*, 2019). Preliminary experiments found that this simple approach performs equally well compared to other more sophisticated methods, e.g. sample similar images that have nearby labels etc, as long as the batch size is large enough, e.g. 100.

3.3 Efficient Indexing and Inference

One major advantage of VisualSparta is how one can use it for real-time inference. That is for a testing query $q = [w_0, ...w_m]$, the ranking score between q and an image is:

$$\text{CACHE}(w, v) = \log(\text{Eq. 11}) \quad w \in W \tag{14}$$

$$f(q, v) = \sum_{i=1}^{m} \text{CACHE}(w_i, v) \tag{15}$$

Since the query term embedding is non-contextual, we can compute the rank feature $\phi(w, v)$ for every possible term w in the vocabulary W with every image v. The resulting score is cached during indexing as shown in Eq. 14. At inference time, the final ranking score can be computed via O(1) look up plus a simple summation as shown in Eq. 15.

More importantly, the above computation can be efficiently implemented via a Inverted Index (Manning *et al.*, 2008), which is the underlying data structure for modern search engines as shown in Figure 1. This enables us to index a very large image dataset using modern search engine, e.g. Elasticsearch (Gheorghe *et al.*, 2015).

4 Experiments

4.1 Datasets

In this paper, we use MSCOCO (Lin *et al.*, 2014) and Flickr30K (Plummer *et al.*, 2015) datasets for the training and evaluation of text-to-image retrieval task. MSCOCO is a large-scale multi-task datasets including object detection, semantic segmentation, and image captioning data. In this experiment, we use the image cpationing dataset split as the source of data for text-to-image model training and evaluation. Following the experimental settings from Karpathy and Fei-Fei (2015), we split the data into 113,287 images for training, 5,000 images for development, and 5,000 images for testing. Each image is paired with 5 captions from different annotators. The performance of both 1,000 and 5,000 test splits are reported and compared with previous results.

Flickr30K (Plummer *et al.*, 2015) is another large scale image captioning datasets, which contains 31, 783 images in total. Following the split from Karpathy and Fei-Fei (2015), 29, 783 images are used for training and 1,000 images are used for testing. The final model is tested on 1,000 test images.

For large-scale efficiency experiment, since there is no existing large-scale image captioning datasets available, we manually design 113K and 1M datasets for testing the inference speed of different models in the large-scale setting. Note that for these two datasets, since we are only interested in speed comparison, the accuracy/quality of the data itself can be ignored. The 113K dataset refers to the MSCOCO training set, which contains 113, 287 images, $\sim$23 times bigger than the MSCOCO 5K test set. The 1M dataset we design consists of 1 million images randomly sampled from the MSCOCO training set. All the efficiency test experiment all done using original MSCOCO 1K and 5K test splits, plus 113K and 1M splits as test beds.

4.2 Evaluation Metrics

Following previous works, we use recall rate as our accuracy evaluation metrics. In both MSCOCO and Flikr30k datasets, we report Recall@k, k=[1, 5, 10] and compare with previous works.

For large-scale efficiency evaluation, we choose query per second and latency(ms) as the evaluation metric to test how each model performs in terms of speed under different sizes of image index.

4.3 Experimental Results

4.3.1 Recall Performance

We compare the model performance with current state-of-the-art retrieval model in text-to-image search. As shown in table 1, the results reveals that our model is competitive compared with previous methods, and achieves state-of-the-art results in most of the evaluation metrics.

Specifically, in MSCOCO 1K test set, our model improves the R@1 performance by 1.9% from previous best methods (Wang *et al.*, 2020), and gets the same results as previous best method in R@5 and R@10. In 5K split, for R@1, 5, 10, our model outperforms the previous best method by 5.8%, 3.5%, and 2.0% respectively. Speaking of Flickr30K dataset, the VisualSparta performance is not higher compared with CVSE (Wang *et al.*, 2020) in terms of for R@5 and R@10, whereas in terms of R@1, VisualSparta is still 1.3% higher than CVSE. In short, VisualSparta achieves best results over previous methods on 7 out of 9 evaluation metrics.

4.3.2 Speed Performance

As discussed in section 4.2, to show the efficiency of VisualSparta model, in addition to original 1K and 5K test split, we also create 113K dataset and 1M dataset, two large-scale benchmark datasets for retrieval speed comparison.

To make fair comparison, we use the optimized hardware for both previous method (GPU accelerated) and our VisualSparta method (CPU-multithread accelerated). For VisualSparta, we use the top-1000 term scores settings for the experiment. For all three models, we use the MSCOCO test-1K split as the source of query to test the speed performance. Since all three models need the same

	MSCOCO-1k			MSCOCO-5k			Flickr 30K		
Model	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SM-LSTM (Huang *et al.*, 2017)	40.7	75.8	87.4	-	-	-	30.2	60.4	72.3
DAN (Nam *et al.*, 2017)	39.4	69.2	79.1	-	-	-	39.4	69.2	79.1
VSE++ (Faghri *et al.*, 2017)	52.0	-	92.0	30.3	-	72.4	39.6	-	79.5
CAMP (Wang *et al.*, 2019b)	58.5	87.9	95.0	39.0	68.9	80.2	51.5	77.1	85.3
SCAN (Lee *et al.*, 2018)	58.8	88.4	94.8	38.6	69.3	80.4	48.6	77.7	85.2
PFAN (Wang *et al.*, 2019a)	61.6	89.6	95.2	-	-	-	50.4	78.7	86.1
CVSE (Wang *et al.*, 2020)	66.3	91.8	96.3	-	-	-	56.1	**83.2**	**90.0**
VisualSparta (ours)	**68.2**	**91.8**	**96.3**	**44.4**	**72.8**	**82.4**	**57.4**	82.0	88.1

Table 1: Detailed comparisons of text-to-image retrieval results in MSCOCO (1K/5K) and Flickr30K datasets

Index Size vs. Query/s	OSCAR	CVSE	Visual Sparta
1K	0.4	177.4	451.4
5K	0.06	162.0	390.5
113K	0.003	5.4	275.5
1M	0.0003	0.3	117.3

Table 2: **Model Speed vs. Index Size**: VisualSparta experiments are done under setting top-K term scores to 1000. All experiments are performed using default acceleration hardware as described in 4.3.2.

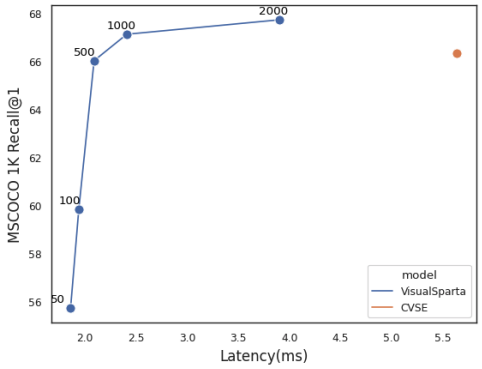

Figure 2: **Inference Speed vs. Model Accuracy.** Each dot represents model performance using top-K term scores. Larger K gives higher accuracy and longer latency time. By setting K to 1000, our model already outperforms CVSE model, with ∼2.3X speedup. Experiments are done on MSCOCO 1K test split.

form of Faster-RCNN image region features as input, we does not take the processing time of this part into consideration. Instead, image region features are directly used as input.

As we can see from Table 2, in all four data split, VisualSparta model significantly outperforms the best dual-encoders retrieval model (CVSE (Wang *et al.*, 2020)) and the best transformer-based retrieval model (Oscar (Li *et al.*, 2020)). Specifically, in 113K dataset, the speed of VisualSparta is 51 times faster than CVSE model, and 91,833 times faster than Oscar model. In 1M dataset, the speed of VisualSparta is 391 times faster than CVSE model. and 391, 000 times faster than Oscar model.

Table 2 also reveals that as the number of images increases, the performance drop is much slower when comparing VisualSparta with other two methods. When the dataset size increases from 113K to 1M, the speed of VisualSparta only decreases by 2.35 times. Compared with it, the speed of CVSE decreases by 13.5 times, whereas the speed of Oscar decreases by 7 times, both of which substantially higher than that of VisualSparta. This experiment shows the absolute speed advantage of VisualSparta compared with previous methods in the large-scale settings.

4.4 Implementation Details

All experiments are done using PyTorch library. During training, one NVIDIA Titan X GPU is used. During speed performance evaluation, for CVSE and Oscar model, one NVIDIA Titan X GPU is used for acceleration. For VisualSparta, a 10-core Intel 9820X CPU is used. For the image encoder, we initialize the model weights from Oscar-base model (Li *et al.*, 2020) with 12 layers and 768 hidden dimensions. For the query embedding, we also initialize it from the Oscar-base word embedding. The learning rate is set to 5e-5 with batch size 160. The number of training epochs is set to 20.

5 Model Analysis

5.1 Speed-Accuracy Flexibility

As described in 3.3, each image can be well represented by a list of weighted tokens independently. This feature makes VisualSparta flexible during indexing time: users can choose to index using top-

Figure 3: Examples of retrieved images and corresponding sparse embedding (terms) under MSCOCO 113K split.

			MSCOCO-1k			MSCOCO-5k		
K	latency (ms)↓	query/s↑	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
50	1.9	537.0	55.7	83.8	91.3	32.5	59.8	71.1
100	1.9	514.7	59.8	87.0	93.6	36.6	64.5	75.3
500	2.1	477.7	66.0	90.2	95.5	41.9	70.4	80.2
1000	2.4	414.5	67.1	91.0	95.8	43.3	71.5	81.4
2000	3.9	256.3	67.7	91.5	96.2	44.2	72.4	82.0
all	6.9	144.1	68.2	91.8	96.3	44.4	72.8	82.4

Table 3: Effect of top-K term scores in terms of speed and accuracy tested in MSCOCO dataset; ↑ means higher the better, and ↓ means lower the better

K term scores based on their memory constraint or speed requirement. Table 3 compares different choices of K value and their corresponding recall and speed performance, in both MSCOCO 1K and 5K split. Figure 2 visualizes the trade-off between model accuracy and inference speeed. The x axis represents the latency(ms) of a single query, and the y axis denotes the Recall@1 score under MSCOCO 1K test set. Each dot represents the model performance under certain top-K term score settings. The curve reveals that with the increase of the K, the recall becomes higher, whereas each query takes longer time to get the retrieval result. From the comparison between VisualSparta and CVSE model, we observe that VisualSparta can already beat the accuracy performance of CVSE when setting top-K term score to 1000. In this small-scale case with only 1K image index, we already get an absolute ∼2.3X speedup.

5.2 Retrieved Images in a Large-scale Setting

Since no public large-scale retrieval dataset existed, we query the VisualSparta model on the 113K split and manually check the results. As shown in Figure 3, most of the retrieved results make sense and are highly relevant to the query provided. This shows that VisualSparta is capable of retrieving relevant results in the large-scale settings. Moreover, by inspecting top terms corresponding to each retrieved image, we observe that these terms are also very relevant to their corresponding images, which implies that weighted terms is a valid and rich representations for these images.

6 Conclusion

In conclusion, this paper presents VisualSparta, an accurate and efficient text-to-image retrieval model which shows the state-of-the-art scalable performance in both MSCOCO and Flickr30K. Its main novelty lies in the combination of powerful pre-trained image encoder with fragment-level scoring. Detailed analysis also demonstrates that our approach has substantial scalability advantages compared to previous best methods when indexing large image datasets for real-time searching, making it suitable for real-world deployment.

References

Zhuyun Dai and Jamie Callan. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv preprint arXiv:1910.10687*, 2019.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

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.

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.

Radu Gheorghe, Matthew Lee Hinman, and Roy Russo. *Elasticsearch in action*. Manning, 2015.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. Convert: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688*, 2019.

Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318, 2017.

Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.

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.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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.

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.

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 *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. Expansion via prediction of importance with contextualization. *arXiv preprint arXiv:2004.14245*, 2020.

Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

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.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748*, 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 5764–5773, 2019.

Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *European Conference on Computer Vision*, pages 18–34. Springer, 2020.

Jonatas Wehrmann, Douglas M Souza, Mauricio A Lopes, and Rodrigo C Barros. Language-agnostic visual-semantic embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5804–5813, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. Sparta: Efficient open-domain question answering via sparse transformer matching retrieval. *arXiv preprint arXiv:2009.13013*, 2020.