Relational Visual-Textual Information Retrieval

Nicola Messina[0000−0003−3011−2487]

Institute of Information Science and Technologies, National Research Council, Pisa, Italy
nicola.messina@isti.cnr.it

Abstract. With the advent of deep learning, multimedia information processing gained a huge boost, and astonishing results have been observed on a multitude of interesting visual-textual tasks. Relation networks paved the way towards an attentive processing methodology that considers images and texts as sets of basic interconnected elements (regions and words). These winning ideas recently helped to reach the state-of-the-art on the image-text matching task. Cross-media information retrieval has been proposed as a benchmark to test the capabilities of the proposed networks to match complex multi-modal concepts in the same common space. Modern deep-learning powered networks are complex and almost all of them cannot provide concise multi-modal descriptions that can be used in fast multi-modal search engines. In fact, the latest image-sentence matching networks use cross-attention and early-fusion approaches, which force all the elements of the database to be considered at query time. In this work, I will try to lay down some ideas to bridge the gap between the effectiveness of modern deep-learning multi-modal matching architectures and their efficiency, as far as fast and scalable visual-textual information retrieval is concerned.

Keywords: Cross-media Retrieval · Deep Features · Neural Networks

1 Introduction

Image-text matching has shown impressive results on many image-sentence retrieval benchmarks, where the objective consists in retrieving images given a sentence as a query, or vice versa. The image-sentence retrieval task has been used to evaluate the network's ability to correctly match together relevant images and sentences. However, the image-sentence retrieval problem is interesting in itself, as it lays the basis for efficient search engines working with multi-modal data.

Search engines must be fast and scalable, as they need to process queries on huge databases in few milliseconds. However, state-of-the-art image-sentence matching approaches usually employ cross-attention mechanisms in the early stages of the data pipeline that makes it impossible to separately forward the visual and the textual information (Figure 1). This separation is needed to disentangle the offline indexing phase, usually very expensive, from the online query processing, that instead should be completed in milliseconds.

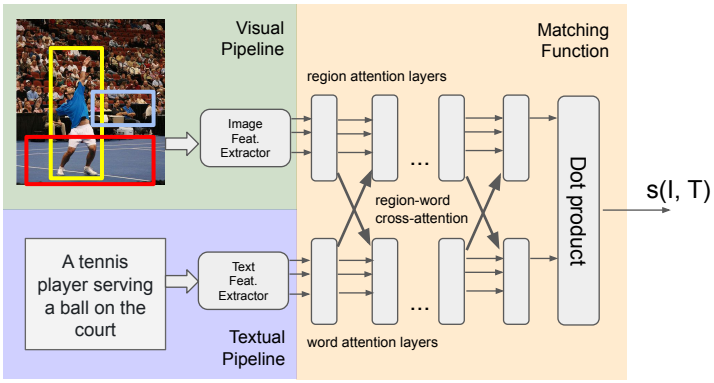

Fig. 1: The overall architecture of state-of-the-art proposals concerning image-text matching. Cross attention in early stages makes the matching function outputting image-text similarity score very complex. It is not possible to extract separately visual and textual features.

Despite the overall loss in efficiency, the use of cross-attention produces an effective multi-step reasoning process that is highly beneficial for producing good matches. As shown in previous works such as Relation Networks [16], trying to infer a relational bias between the basic building blocks of the visual and textual inputs helps in developing abstract links between multi-modal concepts to gather a relational view of the world.

Furthermore, it is possible that the optimal image-sentence representation for good indexing is not a fixed-sized vector, but a variable-length set of vectors describing the images and the texts as sets of concepts. This poses new challenges as far as the indexing structures are concerned.

In this work, I will try to pave the way towards the use of effective relational multi-modal descriptions obtained from state-of-the-art self-attentive architectures in scalable retrieval contexts, where efficiency is a key requirement.

2 Related Work

Many works in computer vision and natural language processing works introduced high-level complex reasoning mechanisms [16,19], mainly addressing Visual Question Answering. More recently, the basic ideas behind these reasoning schemes have been implemented in self- and cross-attentive modules [18], and employed in many language-vision tasks [3,7,15]. These works achieved state-of-the-art results on image-sentence retrieval. However, they do not consider efficiency aspects.

On the other hand, many works tackled the problem of indexing visual features coming from deep architectures. As far as content-based image retrieval (CBIR) is concerned, [2,1] addressed the indexability problem of deep visual

features coming from Convolutional Neural Networks (CNNs), like [17]. In particular, [1] showed the performances achieved with the quantization of RMAC features, and they compared this methodology to the deep permutation approach [2], using an inverted file as an indexing structure. Standard CNN features do not embed complex relational biases. Furthermore, multi-modality is not addressed in these works.

3 Explored Approaches

In previous publications [10,11,12] I explored the effectiveness of a relational visual descriptor extracted from a relation-aware architecture reasoning on a scene with multiple objects. This relational descriptor obtained best results in the introduced Relational-CBIR task, which consists in finding all the images having objects in similar spatial relationships, given an image as query. The proposed relational feature defeated common CBIR deep features such as RMAC [17] on this task. This work tackled the problem of producing a compact and effective visual descriptor that could carry very complex scene information, including inter-object relationships, and that could be indexed using already-existing CBIR frameworks.

Given the increasing interest in multi-modal relational information processing, my research is now focused on complex cross-modal retrieval scenarios. The excellent results obtained by recently introduced self- and cross-attentive models made me concentrate on the transformer architecture [18] for processing visual-textual data using multi-step reasoning pipelines.

Although current efficient retrieval methods assume fixed-sized descriptors (e.g., RMAC [17]), the latest works in cross-modal analysis treat images and texts as sets of basic interconnected elements (image regions and words) processed using attentive mechanisms. The native representation available becomes therefore a variable set of features, called *concepts*, for every image or sentence.

One of my recent key contributions in this direction is the introduction of the Transformer Encoder Reasoning Network (TERN) [14]. TERN employs self-attentive mechanisms to produce both a global fixed-size deep feature and sets of fine-grained concepts that are independent of their source modality. Unlike most works in this field, TERN lacks cross-attentive links. Doing so, two well-distinguished pipelines, a visual and a textual one, are created and can be used separately in the online search and in the offline indexing phases. The produced cross-modal representations are compared with simple dot-products so that the similarity search can be very efficient by employing already existing indexing schemes working on standard metric spaces (Figure 2).

Concerning the global fixed-size description of images and sentences, in [13] we applied the scalar quantization or deep permutation approaches to multi-modal global features as explained in [1]. These representations can be then easily used in inverted lists without further modifications to the indexing structure.

On the other hand, TERN can also treat images and sentences as sets and sequences of basic interconnected concepts, coming from regions and words re-

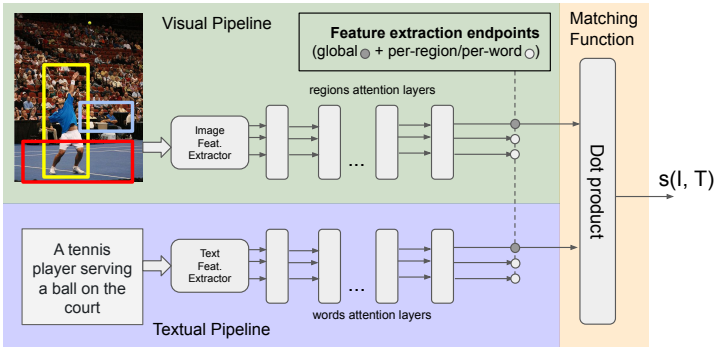

Fig. 2: An high-level overview of the proposed TERN architecture.

spectively. In this case, TERN does not output a compact global description of images and sentences, but a variable-sized set of features, one for every concept, in the same abstract common space.

The concepts can be clustered to create a dictionary. Following this direction, in [13] we also introduced a model similar to the Bag of Words, that we called Bag of Concepts, for producing image and sentence representations for efficient indexing using inverted lists.

3.1 Early Results

In [11] we were able to obtain a good relation aware descriptor, that reached a Spearman-Rho correlation value of 0.28 against -0.15 of the RMAC features on the Relational-CBIR benchmark built on the CLEVR dataset [5]. We thus showed the efficacy of a relational architecture in producing a fixed-size relation aware image descriptor.

In the recent work on visual-textual information retrieval [13] we used the proposed TERN architecture [14] as a multi-modal feature extractor, both for global fixed-sized descriptors and for the variable-sized set of concepts. In these first experiments, we tested the stability of the extracted features by simulating strong sparsification, as this is the key element for the production of efficient inverted indexes.

On the visual-textual MS-COCO dataset [6], the scalar quantization and the deep permutation approaches on the fixed-sized global feature behaved very similarly (72.7 Recall@10 in the sentence-to-image retrieval and 81.3 in image-to-sentence). When the sparsity rate achieves 99% (only 20 dimensions out of 2048 are not zero), the deep permutation approach loses around 27% on the Recall@10 metric, while scalar quantization loses 23%. At the same very high sparsification rate of 99%, the Bag of Concepts model shows better results than the scalar quantization approach during the re-ranking phase of the sentence-to-image retrieval scenario. The reranking using the non sparsified vectors can be performed efficiently using GPUs on the subset of results selected by the initial

approximate search, therefore it defines an overall good compromise between efficiency and effectiveness.

4 Conclusions

In this work, I tried to pave the way towards efficient and effective multi-modal retrieval using state-of-the-art technologies from computer vision and natural language processing worlds. The emphasis is placed on attentive architectures. They are able to implement a multi-step high-level relational reasoning procedure, gaining a lot in effectiveness but creating efficiency problems when scalable information retrieval is addressed.

In my research, I first tried to produce a fixed-sized relational visual descriptor that defeated RMAC on the Relational-CBIR task. Then, considering the interesting cross-modal retrieval problem, I tried to extract powerful fixed- and variable-sized features using the proposed TERN architecture, using existing methods (scalar quantization, deep permutations) when addressing global fixed-sized features and proposing the Bag of Concepts model for producing indexable representations out of the variable-sized sets of multi-modal concepts.

4.1 Next steps

In the near future, I manage to extensively evaluate the efficiency of the proposed approaches when implemented in inverted indexes structures. To do so, a large multi-modal dataset containing matching images and sentences is needed. MS-COCO can be augmented with Flickr30k [20], obtaining a total of around 36k images annotated with 180k sentences. For further validating these approaches, a whole set of distracting images can be added from available huge image datasets such as MIRFlickr1M [1].

Further experimentation is needed as far as the Bag of Concepts is concerned. An extension of the TERN architecture that I am implementing involves the fine-grained alignment of regions and words at training time. In this case, a precise similarity matrix between every image region and every word is available. It is therefore possible that a custom indexing structure can be built using the region-word alignment matrix. It is furthermore possible to learn this indexing structure imposing some sparsification constraints directly at training time.

Another line of research can be derived by the approaches by [8,9]. In these works, the transformer encoder in the BERT architecture [4] is split in an offline and an online processing stages by partitioning the attention links so that they do not create cross-connections between the two pipelines. In this way, the complex activations produced from the offline pipeline can be stored during the indexing phase and efficiently retrieved at query time. They applied this methodology for textual document retrieval. However, this approach can be directly applied to state-of-the-art visual-textual processing architectures based on the BERT model, such as [3,7,15].

[1] bluehttp://press.liacs.nl/mirflickr/

References

1. Amato, G., Carrara, F., Falchi, F., Gennaro, C., Vadicamo, L.: Large-scale instance-level image retrieval. Information Processing & Management p. 102100 (2019)
2. Amato, G., Falchi, F., Gennaro, C., Vadicamo, L.: Deep permutations: Deep convolutional neural networks and permutation-based indexing. In: Amsaleg, L., Houle, M.E., Schubert, E. (eds.) SISAP 2016. Lecture Notes in Computer Science, vol. 9939, pp. 93–106 (2016)
3. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740 (2019)
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT 2019. pp. 4171–4186. Association for Computational Linguistics (2019)
5. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2901–2910 (2017)
6. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV 2014. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014)
7. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS 2019. pp. 13–23 (2019)
8. MacAvaney, S., Nardini, F.M., Perego, R., Tonellotto, N., Goharian, N., Frieder, O.: Efficient document re-ranking for transformers by precomputing term representations. arXiv preprint arXiv:2004.14255 (2020)
9. MacAvaney, S., Nardini, F.M., Perego, R., Tonellotto, N., Goharian, N., Frieder, O.: Expansion via prediction of importance with contextualization. arXiv preprint arXiv:2004.14245 (2020)
10. Messina, N., Amato, G., Carrara, F., Falchi, F., Gennaro, C.: Learning relationship-aware visual features. In: Leal-Taixé, L., Roth, S. (eds.) Computer Vision – ECCV 2018 Workshops. pp. 486–501. Springer International Publishing, Cham (2019)
11. Messina, N., Amato, G., Carrara, F., Falchi, F., Gennaro, C.: Learning visual features for relational cbir. International Journal of Multimedia Information Retrieval (Sep 2019). https://doi.org/10.1007/s13735-019-00178-7, bluehttps://doi.org/10.1007/s13735-019-00178-7
12. Messina, N., Amato, G., Falchi, F.: Re-implementing and extending relation network for r-cbir. In: Ceci, M., Ferilli, S., Poggi, A. (eds.) Digital Libraries: The Era of Big Data and Data Science. pp. 82–92. Springer International Publishing, Cham (2020)
13. Messina, N., Amato, G., Falchi, F., Gennaro, C., Marchand-Maillet, S.: Cross-media visual and textual retrieval using transformer-encoder deep features. SISAP 2020 (Submitted)
14. Messina, N., Falchi, F., Esuli, A., Amato, G.: Transformer reasoning network for image-text matching and retrieval. arXiv preprint arXiv:2004.09144 (2020)
15. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. CoRR **abs/2001.07966** (2020)

16. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: Advances in neural information processing systems. pp. 4967–4976 (2017)
17. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS 2017. pp. 5998–6008 (2017)
19. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–29 (2016)
20. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)