Image Re-ranking with Long-Context Sequence Modeling

Zilin Xiao, Ayush Sachdeva, Hao-Jen Wang, and Vicente Ordonez

Department of Computer Science, Rice University, USA
`{zilin, as216, dw63, vicenteor}@rice.edu`

Abstract. We introduce ExtReranker, a model that takes as input local features corresponding to an image query and a group of gallery images, and outputs a refined ranking list through a single forward pass. ExtReranker formulates the **re-rank**ing problem as a span **ext**raction task analogous to the text span extraction problem in natural language processing. In contrast to pair-wise correspondence learning, our approach leverages long-context sequence models to effectively capture the list-wise dependencies between query and gallery images at the local-feature level. ExtReranker achieves state-of-the-art re-ranking performance compared to alternative methods on $\mathcal{R}$Oxford and $\mathcal{R}$Paris while using 10× fewer local descriptors and having 5× lower forward latency.

1 Introduction

Instance-level image retrieval is an important problem in computer vision with many applications. In general, retrieval is usually cast as a metric learning problem where a model is trained under a distance or similarity objective to compare pairs of inputs. Due to the high dimensional nature of images, this process is typically accomplished in two stages: First, images are mapped to a compact feature representation that can be used with a similarity function for fast retrieval of a candidate gallery set of potential matches. Subsequently, a more powerful but often more computationally demanding re-ranking model refines the retrieved gallery set into a more precise ranked list.

Prior image re-ranking methods [10, 21, 23, 28] typically adopt a pair-wise training objective for scoring positive and negative image pairs accordingly. This strategy does not model nuanced and more complex relative differences from images in the gallery set, *i.e.* the top-scoring image should also influence the relative ranking of other images in the gallery.

In this work, we present ExtReranker, a new image re-ranker that learns from list-wise re-ranking supervision. ExtReranker jointly considers dependencies across multiple candidate images given a query image so that these candidates can be implicitly modeled to calibrate the relevance scores. Figure 1 shows an overview of how ExtReranker compares to pair-wise re-rankers by considering the gallery set jointly. For instance, consider the example provided in this figure, where the top-scored image clearly matches some features at the

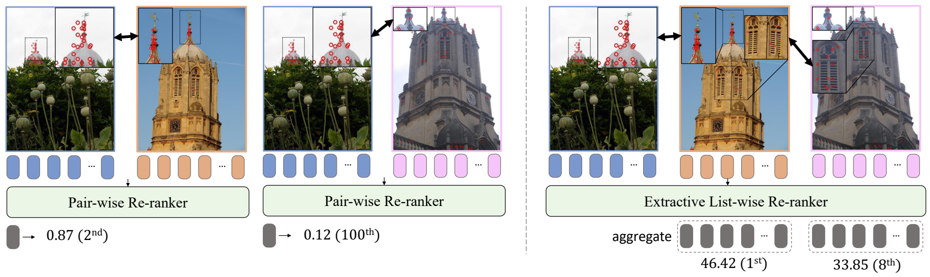

Fig. 1: Overview of pair-wise vs our proposed list-wise re-ranking. Red circles denote the locations of input local features. **Left:** The pair-wise re-ranker gets a high score for a positive image since it clearly depicts the same structure at the top of the tower as the query, while a different positive image gets a low score because the top of the tower is not as clearly visible. **Right:** Our extractive re-ranker can output a high score for both positive image results since it can exploit the transitive relationship between these images as the two gallery images also share common local features.

top of the tower. The top of the tower, on the other hand, is not clearly visible in the second image, which is still a picture of the same tower. However, the list-wise re-ranker can still rank with a high score for this second image due to the transitive relationship between the query image and these two images in the gallery set.

ExtReranker takes inspiration from natural language processing tasks such as sequence tagging [8, 13, 19] and extractive question-answering [5, 18, 20, 27]. Modern solutions to these problems often involve a sequence model that predicts token-level scores to extract task-specific text spans [2, 7]. We adopt this strategy coupled with a long-context transformer model to be able to input local features from both the query image and the entire set of gallery images. Our models trained on Google Landmarks v2 (GLDv2) [25] are state-of-the-art re-rankers in relative performance gains on the Revisited Oxford and Paris datasets [14–16], while relying on significantly fewer local descriptors per image and achieving a 5× decrease in inference latency.

2 Related Works

Given a query image, the goal of image retrieval is to search for similar images in an image database. Early works use hand-crafted local features for image retrieval [3, 11]. Later works divide the image retrieval into a global retrieval stage, where retrieving images using a global descriptor [9, 24], and a subsequent local re-ranking stage [1], where the top-k retrieved images are re-ranked through local feature matching with RANSAC [6]. With advancements in deep learning, global and local features extracted from neural models [4, 10, 12, 26, 29] have replaced handcrafted features. More recently, researchers have attempted to use sophisticated pooling [17] and nearest-neighbor expansion techniques to re-frame

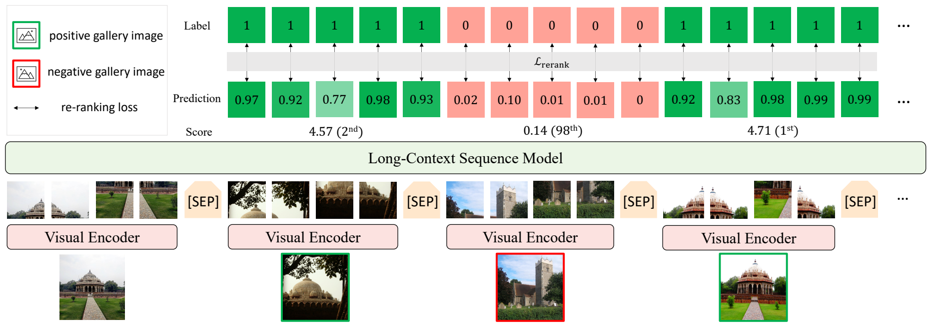

Fig. 2: Training and inference of ExtReranker. At training time, the model optimizes a binary cross-entropy loss on each gallery image token. At inference time, the token scores of each gallery image get aggregated to obtain a re-ranked image list.

image re-ranking to solely consider global image features [21] or even seek a unified single-stage image retrieval solution [22, 30, 31].

3 Methodology and Experiments

Given a query image $\mathbf{I}_q$ and a list of k gallery images $\mathbf{I}_{g,i}, i \in \{1, \cdots, k\}$ returned by global similarity search, the purpose of image re-ranking is to produce another refined list of gallery images that are reordered based on improved similarity measures to the query image $\mathbf{I}_q$. Let $\mathbf{x}_q = \left\{\mathbf{x}_{q,j} \in \mathbb{R}^{d_l}\right\}_{j=1}^{L}$ denote L local descriptors (size d_l) of the query image $\mathbf{I}_q$ extracted from a visual backbone and $\mathbf{x}_{g,i} = \left\{\mathbf{x}_{g,i,j} \in \mathbb{R}^{d_l}\right\}_{j=1}^{L}$ denote local descriptors from the i-th gallery image.

Pair-wise re-ranker. A typical neural re-ranker f_ϕ computes a pair-wise confidence score S for each pair of images $(\mathbf{I}_q, \mathbf{I}_{g,i})$:

$$S(\mathbf{I}_q, \mathbf{I}_{g,i}) = f_\phi(\mathbf{x}_{q,1}, \cdots, \mathbf{x}_{q,L}, \mathbf{x}_{g,i,1}, \cdots, \mathbf{x}_{g,i,L}),$$

where f_ϕ optimizes a binary classification loss. The re-ranked list is obtained at test time by sorting confidence scores for all gallery images in descending order.

Extractive list-wise re-ranker. Figure 2 presents the overview of our method. In contrast to the pair-wise re-ranker, ExtReranker constructs an input token sequence that accommodates the query image local descriptors along with the ones of all gallery images:

$$\mathbf{X}(\mathbf{I}_q, \{\mathbf{I}_{g,i}\}_{i=1}^{k}) := [\mathbf{x}_{q,1}; \cdots; \mathbf{x}_{q,L}; [\text{SEP}]; \mathbf{x}_{g,1,1}; \cdots; \mathbf{x}_{g,1,L}; [\text{SEP}]; \cdots; \mathbf{x}_{g,k,1}; \cdots; \mathbf{x}_{g,k,L}; [\text{SEP}]],$$

where [SEP] represents a special token embedding interleaved between descriptors of different images. Let g_φ be a sequence model that provides contextualized representations for each input token. ExtReranker first uses g_φ to produce token-wise representations:

Method	# local desc.	Medium				Hard			
		$\mathcal{R}$Oxf	$\mathcal{R}$Oxf+1M	$\mathcal{R}$Par	$\mathcal{R}$Par+1M	$\mathcal{R}$Oxf	$\mathcal{R}$Oxf+1M	$\mathcal{R}$Par	$\mathcal{R}$Par+1M
RN50-DELG [4]	-	73.6	60.6	85.7	68.6	51.0	32.7	71.5	44.4
+ GV Rerank	1,000	78.3	67.2	85.7	69.6	57.9	43.6	71.0	45.7
+ RRT Rerank [23]	500	78.1	67.0	86.7	69.8	60.2	44.1	75.1	49.4
+ CVNet Rerank [10]	3,072	78.7	67.7	87.9	72.3	63.0	46.1	76.8	52.5
+ ExtReranker-small	50	80.9	70.9	**89.3**	77.7	63.3	49.3	**77.8**	57.9
+ ExtReranker-base	50	**81.6**	**71.7**	89.2	**77.9**	**64.2**	**50.5**	77.6	**58.2**
RN101-DELG	-	76.3	63.7	86.6	70.6	55.6	37.5	72.4	46.9
+ GV Rerank	1,000	81.2	69.1	87.2	71.5	64.0	47.5	72.8	48.7
+ RRT Rerank	500	79.9	-	87.6	-	64.1	-	76.1	-
+ ExtReranker-small	50	81.8	74.1	87.9	75.9	64.2	54.7	75.0	55.2
+ ExtReranker-base	50	**83.3**	**76.1**	**90.3**	**80.7**	**66.9**	**56.2**	**81.4**	**61.8**

Table 1: Image retrieval performance (% mAP) on $\mathcal{R}$Oxf and $\mathcal{R}$Par and their 1M distractor variants (+1M) based on DELG [4] local features with Medium and Hard evaluation strategy. For a fair comparison, results for re-rankers are reported with their top-100 candidates unless indicated otherwise.

$$\begin{aligned}\mathbf{H}(\mathbf{I}_q, \{\mathbf{I}_{g,i}\}_{i=1}^{k}) &= g_\varphi(\mathbf{x}_{q,1}; \cdots; \mathbf{x}_{q,L}; [\mathrm{SEP}]; \mathbf{x}_{g,1,1}; \cdots; \mathbf{x}_{g,1,L}; [\mathrm{SEP}]; \cdots; \mathbf{x}_{g,k,1}; \cdots; \mathbf{x}_{g,k,L}; [\mathrm{SEP}]) \\ &= \left[\mathbf{h}_{q,1}, \cdots, \mathbf{h}_{q,L}, \mathbf{h}^{q}_{[\mathrm{SEP}]}, \mathbf{h}_{g,1,1}, \cdots, \mathbf{h}_{g,1,L}, \mathbf{h}^{g,1}_{[\mathrm{SEP}]}, \cdots, \mathbf{h}_{g,k,1}, \cdots, \mathbf{h}_{g,k,L}, \mathbf{h}^{g,k}_{[\mathrm{SEP}]}\right],\end{aligned}$$

The re-ranking training is enforced via binary cross-entropy loss on each gallery image token position. At inference time, we collect all token scores of a gallery image and use a certain aggregator function p so that each gallery image gets an aggregated score. Then, the list-wise refined score can be used to construct the refined gallery list. The aggregator function is defined as one of the following: (i) *summation*, (ii) *product*, (iii) *the first token score* or (iv) *the last token score* of a selected span of tokens.

We report ExtReranker performance on $\mathcal{R}$Oxf and $\mathcal{R}$Par in Table 1. Specifically, all models are trained on gallery candidates using the top-50 local descriptors obtained using the DELG feature extractor. We observe that ExtReranker-base exhibits clear and consistent improvement over all local-feature re-ranking baselines. The improvement is most pronounced on $\mathcal{R}$Oxf+1M and $\mathcal{R}$Par+1M in the hard setting. Our model also demonstrates scalability as the overall performance improves with model size from small to base. We refer to model configuration and training recipe in the Appendix.

4 Conclusion

This paper presents ExtReranker, the first image re-ranking framework that leverages list-wise re-ranking supervision at the local feature level. With a long-context sequence model, this approach effectively captures dependencies between the query image and each gallery image in addition to the dependencies amongst the gallery images themselves and implicitly learns to calibrate predictions for a more precise ranking list.

References

1. Avrithis, Y., Tolias, G.: Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. Int. J. Comput. Vis. **107**(1), 1–19 (2014) 2
2. Barba, E., Procopio, L., Navigli, R.: Extend: Extractive entity disambiguation. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. pp. 2478–2488. Association for Computational Linguistics (2022) 2
3. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008) 2
4. Cao, B., Araujo, A., Sim, J.: Unifying deep local and global features for image search. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 726–743. Springer (2020) 2, 4
5. Eberts, M., Ulges, A.: Span-based joint entity and relation extraction with transformer pre-training. In: Giacomo, G.D., Catalá, A., Dilkina, B., Milano, M., Barro, S., Bugarín, A., Lang, J. (eds.) ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020). Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 2006–2013. IOS Press (2020) 2
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**, 381–395 (1981) 2
7. Heinzerling, B., Strube, M.: Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 273–291. Association for Computational Linguistics, Florence, Italy (Jul 2019) 2
8. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. CoRR **abs/1508.01991** (2015), `http://arxiv.org/abs/1508.01991` 2
9. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010. pp. 3304–3311. IEEE Computer Society (2010) 2
10. Lee, S., Seong, H., Lee, S., Kim, E.: Correlation verification for image retrieval. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 5364–5374. IEEE (2022) 1, 2, 4
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004) 2
12. Noh, H., Araújo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. IEEE International Conference on Computer Vision (2016) 2
13. Peters, M.E., Ammar, W., Bhagavatula, C., Power, R.: Semi-supervised sequence tagging with bidirectional language models. In: Barzilay, R., Kan, M. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. pp. 1756–1765. Association for Computational Linguistics (2017) 2

14. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA. IEEE Computer Society (2007) 2
15. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA. IEEE Computer Society (2008) 2
16. Radenovic, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 5706–5715. Computer Vision Foundation / IEEE Computer Society (2018), http://openaccess.thecvf.com/content_cvpr_2018/html/Radenovic_Revisiting_Oxford_and_CVPR_2018_paper.html 2
17. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(7), 1655–1668 (2019) 2
18. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100, 000+ questions for machine comprehension of text. In: Su, J., Carreras, X., Duh, K. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. pp. 2383–2392. The Association for Computational Linguistics (2016) 2
19. Ramshaw, L., Marcus, M.: Text chunking using transformation-based learning. In: Third Workshop on Very Large Corpora (1995), https://aclanthology.org/W95-0107 2
20. Segal, E., Efrat, A., Shoham, M., Globerson, A., Berant, J.: A simple and effective model for answering multi-span questions. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. pp. 3074–3080. Association for Computational Linguistics (2020) 2
21. Shao, S., Chen, K., Karpur, A., Cui, Q., Araujo, A., Cao, B.: Global features are all you need for image retrieval and reranking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11036–11046 (October 2023) 1, 3
22. Song, C.H., Yoon, J., Hwang, T., Choi, S., Gu, Y.H., Avrithis, Y.: On train-test class overlap and detection for image retrieval. arXiv preprint arXiv: 2404.01524 (2024) 3
23. Tan, F., Yuan, J., Ordonez, V.: Instance-level image retrieval using reranking transformers. IEEE International Conference on Computer Vision (2021) 1, 4
24. Tolias, G., Avrithis, Y., Jégou, H.: Image search with selective match kernels: Aggregation across single and multiple images. Int. J. Comput. Vis. **116**(3), 247–261 (2016) 2
25. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2 - a large-scale benchmark for instance-level recognition and retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 2
26. Yang, M., He, D., Fan, M., Shi, B., Xue, X., Li, F., Ding, E., Huang, J.: DOLG: single-stage image retrieval with deep orthogonal fusion of local and global features. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 11752–11761. IEEE (2021) 2

27. Yoon, W., Jackson, R., Lagerberg, A., Kang, J.: Sequence tagging for biomedical extractive question answering. Bioinformatics **38**(15), 3794–3801 (Jun 2022) 2
28. Zhang, H., Chen, X., Jing, H., Zheng, Y., Wu, Y., Jin, C.: ETR: an efficient transformer for re-ranking in visual place recognition. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023. pp. 5654–5663. IEEE 1
29. Zhang, Y., Zhang, T., Chen, D., Wang, Y., Chen, Q., Xie, X., Sun, H., Deng, W., Zhang, Q., Yang, F., Yang, M., Liao, Q., Guo, B.: Irgen: Generative modeling for image retrieval. CoRR **abs/2303.10126** (2023) 2
30. Zhu, S., Yang, L., Chen, C., Shah, M., Shen, X., Wang, H.: R2former: Unified retrieval and reranking transformer for place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19370–19380 (June 2023) 3
31. Zhu, Y., Gao, X., Ke, B., Qiao, R., Sun, X.: Coarse-to-fine: Learning compact discriminative representation for single-stage image retrieval. ICCV (2023) 3