Cross-Resolution deep features based Image Search

Fabio Valerio Massoli ✉[0000−0001−6447−1301], Fabrizio Falchi[0000−0001−6258−5313], Claudio Gennaro[0000−0002−0967−5050], and Giuseppe Amato[0000−0003−0171−4315]

ISTI-CNR, via G. Moruzzi 1, 56124 Pisa, Italy
{fabio.massoli, fabrizio.falchi, claudio.gennaro, giuseppe.amato}@isti.cnr.it

Abstract Deep Learning models proved to be able to generate highly discriminative image descriptors, named deep features, suitable for similarity search tasks such as Person Re-Identification and Image Retrieval. Typically, these models are trained by employing high-resolution datasets, therefore reducing the reliability of the produced representations when low-resolution images are involved. The similarity search task becomes even more challenging in the cross-resolution scenarios, i.e., when a low-resolution query image has to be matched against a database containing descriptors generated from images at different, and usually high, resolutions. To solve this issue, we proposed a deep learning-based approach by which we empowered a ResNet-like architecture to generate resolution-robust deep features. Once trained, our models were able to generate image descriptors less brittle to resolution variations, thus being useful to fulfill a similarity search task in cross-resolution scenarios. To asses their performance, we used synthetic as well as natural low-resolution images. An immediate advantage of our approach is that there is no need for Super-Resolution techniques, thus avoiding the need to synthesize queries at higher resolutions.

Keywords: Cross Resolution · Similarity Search · Deep Convolutional Neural Networks · Image Retrieval

1 Introduction

Content-Based Image Retrieval (CBIR) is one of the most active research fields in the multimedia community [9,20]. Key aspects that greatly affect a CBIR system performance are the quality of the used image descriptors and its ability to scale. Before the advent of Deep Learning (DL) techniques, the Scale-Invariant Feature Transform (SIFT) [16] based methods were among the most frequently used to generate image descriptors. It has only been after the breakthrough in

G. Amato et al. (Eds.): SISAP 2020, LNCS 12440, pp. 352360, 2020.
Final authenticated publication: https://doi.org/10.1007/978-3-030-60936-8_27

2012 [14] that the scientific community has turned its attention towards DL techniques as a possible approach to the Image Retrieval (IR) problem [3,19].

Since then, DL algorithms were employed in a variety of other fields such as object recognition [10], speech recognition [7], natural language processing [8], etc. Among the various architectural designs, Convolutional Neural Networks (CNNs) experienced the greatest success in the field of computer vision-related tasks. These models are largely used to fulfill CBIR tasks, too, thanks to their ability to create image representations, called deep features, that can be employed as global descriptors for similarity searches [2,9]. Despite their success, a well-known problem of CNN models is that the discriminative ability of the extracted features degrades when a model is fed with low-resolution images [22,17,18]. A reasonable explanation for this issue is that the datasets typically used to train the CNNs to contain images predominantly at high resolutions.

To overcome this issue, in this paper, we presented the approach we employed to solve the cross-resolution IR task. Specifically, we experimented with the effectiveness of our method in the scenario of Face Image Retrieval (FIR), which is of particular concern, for example, for surveillance systems [22,6] that rely on probe images extracted from cameras with limited resolution. To conduct our experiments, we leveraged a ResNet-50 architecture [11], equipped with Squeeze-and-Excitation blocks [12], pre-trained on the VGGFace2 dataset [5]. Starting from the state-of-the-art model [5], we fine-tuned it to make its deep features less brittle to resolution variations of the input images.

The remaining part of the paper is organized as follows. In Section 2, we briefly reviewed some related works. In Section 3, we presented the experimental procedure alongside the results. In Section 4, we concluded the paper with a summary of the main results and future perspectives.

2 Related Works

Before the advent of the Machine Learning (ML), CBIR was based on the extraction and use of low-level feature descriptors, such as color and edge features [13] or local features [16,4]. With the advent of DL models, researchers started to use their inner activations, called deep features, as descriptors for the input images. The similarity search was then directly carried out among features employing specific metrics.

Lin et al. [15] proposed a two steps-based framework in which they did a first coarse-level search followed by a fine-level one. For each query image, the authors considered two features vectors: a binary and global ones. The former was employed to perform a quick search in the database, while the latter was used to perform the fine-level search. In Ahmad et al. [1], the authors proposed a bilinear model in which image features were accumulated at various locations and scales using the convolutional activations extracted from different inner layers of a CNN. With such an approach, the authors were able to extract image descriptors with high discriminative power. Tzelepi et al [21] proposed a method to retrain a model to empower it to generate descriptors better suited for CBIR applications. Specifically, they proposed three basic model retraining approaches: Fully Unsupervised Retraining; Retraining with Relevance Information; Relevance Feedback based Retraining.

3 Experiments

Our starting point was the state-of-the-art SeNet-50 architecture [5] trained on the VGGFace2 [5] dataset. Subsequently, we fine-tuned it employing our training procedure, and finally, we tested the performance of the model on the FIR task using the deep features extracted from its penultimate layer.

3.1 Dataset

The VGGFace2 [5] dataset consists of ∼3.31 million images shared among ∼9K identities. It is divided into two splits, one for training and one for test purposes only. The latter contains ∼170K images divided into 500 identities, while all the other images belong to the remaining ∼8K classes available for training. The entire dataset is characterized by a very low label noise and by a high intra-class variance, especially among head poses. These characteristics make it a suitable choice for training DL models on face-related tasks. Despite these qualities, the dataset mainly comprises high-resolution images. Indeed, training set images have an average resolution of ∼137x180 pixels with less than 1% at a resolution below 32 pixels. As it is shown later, this makes the internal representations of a neural network trained on this dataset brittle to resolution variations in the input data.

3.2 Training details

To fine-tune the model, we made a first trial in which we kept the entire net frozen except for the last fully connected layer, and we fed it with low-resolution images. To obtain the desired inputs, we leveraged the bilinear interpolation algorithm, implemented in the PIL python library, to down-sample the images at a specific (low) resolution. However, we obtained better results by fine-tuning the entire neural network. The intuition was that there were patterns in the new low-resolution images that the model needed to adjust for. We initially set the value of the learning rate at $5 \cdot 10^{-4}$ and dropped its value by a factor of 5 every time the loss reached a plateau. We used a batch size of 256, a weight decay of 10^{-5}, and a momentum of 0.9.

By only using low-resolution input images, we noticed that the models lost the ability to recognize images at high resolution. For this reason, we introduced a new hyperparameter to control the probability with which to down-sample an image. In this way, the CNNs were trained on both low and high-resolution images at the same time. More specifically, the algorithm we employed to resize images was based on two random extractions: the first one used to decide whether or not to give the image at full resolution, and the second one utilized to set the final resolution at which the image would then be resized in the [8, 256] pixels range. After this first preprocessing phase, the input images were resized so that the size of the shortest side was 256 pixels, later a random crop was applied to select a 224x224 pixels region which matches the input of the network. We split the training dataset into training and validation sets. Specifically, during the training phase, we employed two versions of the latter to monitor the performance of the model on both low and high-resolution domains.

3.3 Similarity Search

After the training phase, we assessed the models' performance on the FIR task. We trained several models with different values of the hyperparameter that controlled the probability with which an image was down-sampled. Specifically, we considered probabilities of 0.1, 0.3, 0.5, 0.7 and 1.0. For each scenario, we took the best models and compared their performance on the FIR task with the baseline model.

3.4 Experimental Results

In Figure 1, we reported a comparison between the queries results obtained by the original pre-trained model and by our fine-tuned ones. Specifically, the results correspond to the first five images returned when the query was downsampled at resolutions of 16 pixels (shortest side). Clearly, the deep representation produced by our model was much more discriminative compared to the one generated by the original model.

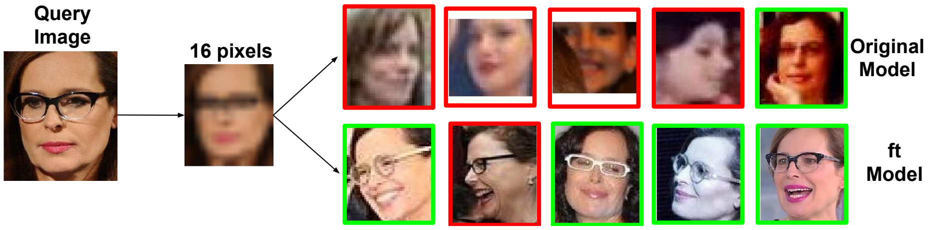

Figure 1. Comparison of the top query results returned by the original and a fine-tuned model considering a query resolution of 16 pixels.

In Figure 2 and Figure 3, we reported the precision and the recall scores, for each fine-tuned model and for the original pre-trained one. Each plot in the figures corresponds to a specific query resolution. As it is clear from Figure 2 and Figure 3, the original pre-trained model experienced a noticeable degradation of its performance when tested against a cross-resolution scenario. With our approach, we have been able to improve upon its performance up to about one order of magnitude, considering queries with resolutions down to 8 pixels, with a negligible loss at higher resolutions.

Finally, in Table 1, we showed the results from the mean Average Precision (mAP) measurements, as a function of the query resolution, for each of the fine-tuned models. Moreover, the first column of the table reported the mAP for the original pre-trained model as a term of comparison. According to Table 1, it is remarkable to notice that our models had higher performance concerning the original pre-trained model in the range between 8 and 24 pixels.

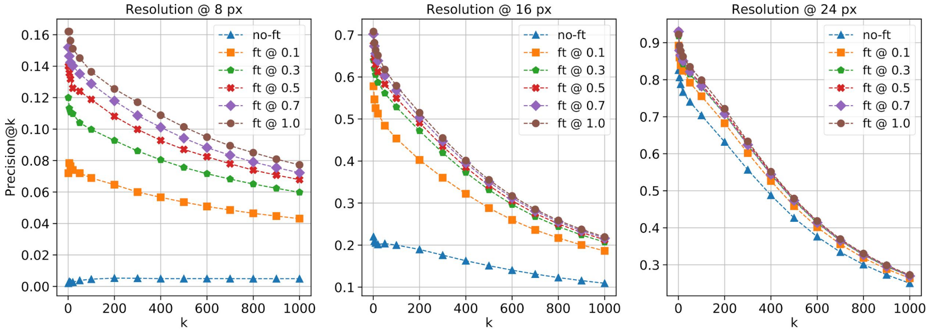

Figure 2. Precision@k. The baseline model has been reported as "no ft" while the "ft" models are the fine-tuned ones. The value after the "@" symbol represents the probability we used during training to decide whether reducing or not the input resolution.

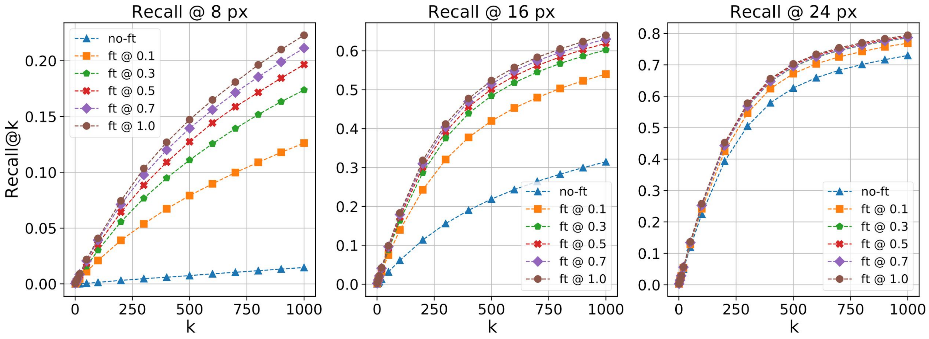

Figure 3. Recall@k. The baseline model has been reported as "no ft" while the "ft" models are the fine-tuned ones. The value after the "@" symbol represents the probability we used during training to decide whether reducing or not the input resolution.

So far, we have considered down-sampled images that have been generated by interpolation methods. Other than that, it is interesting to test our models' performance on datasets composed of native low-resolution images. As a preliminary

Table 1. mAP results, as function of the query resolution, for each model. The baseline model has been reported as “no ft” while the “ft” models are the fine-tuned ones. The value after the “@” symbol represents the probability we used during training to decide whether reducing or not the input resolution.

		no ft	**ft @ 0.1**	**ft @ 0.3**	**ft @ 0.5**	**ft @ 0.7**	**ft @ 1.0**
Query Resolution (pixels)	**8**	0.01	0.05	0.08	0.09	0.09	**0.10**
	16	0.15	0.34	0.40	0.43	0.42	**0.44**
	24	0.56	0.62	0.64	0.65	0.65	**0.66**
	32	0.74	0.75	0.73	0.74	0.75	**0.76**
	64	**0.85**	0.84	0.84	0.84	0.84	0.84
	128	**0.86**	0.85	0.84	0.85	0.84	0.85
	256	**0.86**	0.85	0.84	0.85	0.84	0.85

study, we considered the training set available from the TinyFace dataset[1]. It contains about 8K images with an average height of 20 pixels, distributed among ∼2K different identities. Following the same procedure adopted in the previous measurements, we randomly selected one image for each class as a query, thus extracting the deep features from them and all the other images in the set, to construct the descriptors database. Differently from what was done previously, we did not apply any down-sampling algorithm in this case, since the images were already at low resolution. The precision and recall results were reported in Figure 4, while in Table 2 we reported the mAP values for the fine-tuned models as well as for the original pre-trained model.

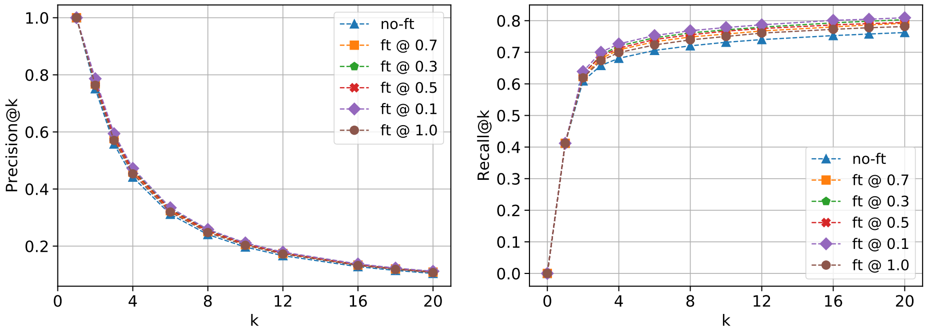

Figure 4. Precision@k (left) and Recall@k (right) for each fine tuned model. The baseline model has been reported as “no ft” while the “ft” models are the fine-tuned ones. The value after the “@” symbol represents the probability we used during training to decide whether reducing or not the input resolution.

[1] https://qmul-tinyface.github.io/

Table 2. mAP results obtained for each model.

no ft	**ft @ 0.1**	**ft @ 0.3**	**ft @ 0.5**	**ft @ 0.7**	**ft @ 1.0**
0.68	0.73	0.72	0.74	0.74	**0.75**

From the results showed in Table 2 it is noticeable that, even though we trained our models on synthetic low-resolution images, their performance was consistent when tested against native low-resolution images and still higher than the original pre-trained model, thus confirming the effectiveness of our training procedure.

4 Conclusion and Future Perspectives

In this paper, we proposed a strategy to train a DL model to generate image descriptors robust against a cross-resolution scenario that can be used to fulfill IR tasks. Specifically, to assess the effectiveness of our method, we considered the task of the FIR, being of particular interest for applications such as surveillance systems. Indeed, in such cases, a probe image that is typically acquired from a security camera has to be matched against a database of known identities, typically characterized by high-resolution image descriptors. Since the security cameras do not usually shoot at very high resolution, and considering that they are often far from the scene, the extracted image can be at resolutions as low as 16 pixels, or even below.

We showed that training a model on a vast dataset, even though it has low noise level and high intra-class variance such as VGGFace1 [5], does not guarantee the robustness of its representations against resolution variations. By using our training method, we were able to improve upon a state-of-the-art CNNs performance on the FIR task, considering a cross-resolution scenario, up to one order of magnitude for query resolutions ranging from 8 to 24 pixels. Besides, we noticed a negligible drop in the performance at resolutions higher than 64 pixels. Therefore, the models trained by embodying our idea were able to produce deep features to be used as global descriptors for images, with sufficient discrimination power among a wide range of resolutions.

Concerning our study, it is clear that the problem of cross-resolution IM is still an open issue. We plan to continue towards this direction by testing new training procedures and by considering new and larger datasets that consist of native low-resolution images to train and test the CNNs in more realistic situations.

Acknowledgments. This research has been funded by the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI) of the National Research Council (CNR). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

1. Alzu'bi, A., Amira, A., Ramzan, N.: Content-based image retrieval with compact deep convolutional features. Neurocomputing **249**, 95–105 (2017)
2. Amato, G., Falchi, F., Gennaro, C., Vadicamo, L.: Deep permutations: deep convolutional neural networks and permutation-based indexing. In: International Conference on Similarity Search and Applications. pp. 93–106. Springer (2016)
3. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: European conference on computer vision. pp. 584–599. Springer (2014)
4. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: European conference on computer vision. pp. 404–417. Springer (2006)
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 67–74. IEEE (2018)
6. Cheng, Z., Zhu, X., Gong, S.: Surveillance face recognition challenge. arXiv preprint arXiv:1804.09691 (2018)
7. Chiu, C.C., Sainath, T.N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R.J., Rao, K., Gonina, E., et al.: State-of-the-art speech recognition with sequence-to-sequence models. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4774–4778. IEEE (2018)
8. Deng, L., Liu, Y.: Deep Learning in Natural Language Processing. Springer (2018)
9. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. pp. 647–655 (2014)
10. Han, J., Zhang, D., Cheng, G., Liu, N., Xu, D.: Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Processing Magazine **35**(1), 84–100 (2018)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
13. Jain, A.K., Vailaya, A.: Image retrieval using color and shape. Pattern recognition **29**(8), 1233–1244 (1996)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
15. Lin, K., Yang, H.F., Hsiao, J.H., Chen, C.S.: Deep learning of binary hash codes for fast image retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 27–35 (2015)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
17. Massoli, F.V., Amato, G., Falchi, F.: Cross-resolution learning for face recognition. Image and Vision Computing p. 103927 (2020)
18. Massoli, F.V., Amato, G., Falchi, F., Gennaro, C., Vairo, C.: Improving multi-scale face recognition using vggface2. In: International Conference on Image Analysis and Processing. pp. 21–29. Springer (2019)
19. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879 (2015)

20. Torresani, L., Szummer, M., Fitzgibbon, A.: Efficient object category recognition using classemes. In: European conference on computer vision. pp. 776–789. Springer (2010)
21. Tzelepi, M., Tefas, A.: Deep convolutional learning for content based image retrieval. Neurocomputing **275**, 2467–2478 (2018)
22. Zou, W.W., Yuen, P.C.: Very low resolution face recognition problem. IEEE Transactions on image processing **21**(1), 327–340 (2011)