Fast and accurate person re-identification with Xception Conv-Net and C2F

Arthur van Rooijen[1,2], Henri Bouma[1], and Fons Verbeek[2]

[1] TNO, Oude Waalsdorperweg 63, 2597 AK The Hague, Netherlands
[2] Leiden University, Niels Bohrweg 1, 2333 CA Leiden, Netherlands

Abstract. Person re-identification (re-id) is the task of identifying a person of interest across disjoint camera views in a multi-camera system. This is a challenging problem due to the different poses, viewpoints and lighting conditions. Deeply learned systems have become prevalent in the person re-identification field as they are capable to deal with the these obstacles. Conv-Net using a coarse-to-fine search framework (Conv-Net+C2F) is such a deeply learned system, which has been developed with both a high-retrieval accuracy as a fast query time in mind. We propose three contributions to improve Conv-Net+C2F: 1) training with an improved optimizer, 2) constructing Conv-Net using a different Convolutional Neural Network (CNN) not yet used for person re-id and 3) coarse descriptors having fewer dimensions for improved speed as well as increased accuracy. With these adaptations Xception Conv-Net+C2F achieves state-of-the-art results on Market-1501 (single-query, 72.4% mAP) and the new, challenging data split of CUHK03 (detected, 42.6% mAP).

Keywords: Person Re-identification, Large-Scale Person Retrieval, Convolutional Neural Networks, Image Retrieval, Feature Extraction

1 Introduction

Person re-identification (re-id) is the task of finding the same person in multiple (surveillance) video resources. This is relevant, because a manual search for individuals in these resources is too laborious and is infeasible for real-world camera networks. In recent years the field of person re-identification has seen an improvement in the accuracy on challenging datasets like Market-1501 [23] (see Table 5). However, most research focuses just on accuracy without taking retrieval time into account. As a consequence, such research is less usable in practice.

Exception to this is the work of Yao et al. [21] who combine a high accuracy with a fast retrieval time. They demonstrate state-of-the-art Rank-1 accuracy of 75.13% on CUHK03, 84.64% (single-query) on Market-1501 and 64.58% (single-query) on the extended version of Market-1501 containing 520k images. For the latter, an average query time of 180ms is obtained. They employ Conv-Net

Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications CIARP-2018, Springer LNCS 11401, pp. 611-619, 2019.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-030-13469-3_71

with a coarse-to-fine retrieval framework (Conv-Net+C2F). The C2F retrieval framework uses coarse descriptors to narrow down the search space, this enables fast searching. Subsequently fine descriptors are used to find matches in the reduced gallery which results in the high accuracy.

We have investigated an improved Conv-Net+C2F which advances the state-of-the-art w.r.t. accuracy and retrieval time on Market-1501 and CUHK03. To this end three modifications to Conv-Net+C2F are proposed: 1) For increased accuracy Adadelta [22] is used instead of Stochastic Gradient Descent [13] (SGD) as optimizer for training Conv-Net. 2) Constructing Conv-Net using Xception [2] instead of GoogLeNet [17] further improves this accuracy. 3) Decreasing the dimensionality of the coarse descriptor to 8 instead of 128 reduces both the average query time and increases the accuracy.

This paper is organized as follows: Related work (Sec. 2), Method (Sec. 3), Experimental setup (Sec. 4), Results (Sec. 5) and Conclusion (Sec. 6).

2 Related work

Our research leads to direct improvement of the Conv-Net with the coarse-to-fine (C2F) retrieval framework proposed by Yao et al. [21]. Having an understanding of this Conv-Net+C2F is therefore crucial. Conv-Net is created by replacing the fully connected layers in GoogLeNet with a classifier block. This classifier block consists of a convolutional layer with C kernels, thereby producing a single feature map of $H \times W$ for every class in the training dataset. The values in each feature map are combined into a single average score per feature map (representing confidence) using Global Average Pooling (GAP). After training the model using SGD, this added classifier block is discarded entirely and the output of the current last convolutional layer of Conv-Net is used to produce image descriptors.

The coarse-to-fine (C2F) retrieval framework enables the fast ranking of the gallery image descriptors w.r.t. a query image descriptor. Distances between descriptors are calculated by taking the Euclidean distance after L2 normalizing each descriptor. The C2F retrieval framework makes a trade-off between accuracy and speed by using two different descriptors: a coarse descriptor and a fine descriptor. The small coarse descriptor is created by applying GAP to the filters in the last convolutional layer of Conv-Net. This results in a K-dimensional vector with K equal to the number of filters in the last layer. This is then reduced to 128 using PCA. The large fine descriptor is created by applying GAP to four equal-sized, non-overlapping, horizontal bands on every filter, thereby creating a $4 \times K$-dimensional fine descriptor. The coarse descriptors are used to rapidly reduce the search space to $M = 500$ images, after which re-ranking is performed using the fine descriptors. Since both the descriptors are computed using the same activation function, the network needs to be used only once in order to construct both. This is described in detail by Yao et al. [21].

3 Proposed approach

We propose three additions to Conv-Net+C2F in order to increase performance speed and accuracy. The first adaptation is to use an improved optimizer for training. Conv-Net is originally trained using SGD, which uses a learning-rate scheme that is experimentally found to work well by Yao et al. [21]. We wish to move away from the manual tuning of the optimizer and therefore suggest the use of the Adadelta [22] optimizer. Besides automatically tuning the learning rate, this also improves the accuracy of the model.

The second improvement to the original Conv-Net is using an improved base-model. A model is transformed into its Conv-Net variant by removing all dense layers present in the network. Next a classifier block is appended to this reduced model during training; similar to the original Conv-Net based on GoogLeNet. In the testing phase this classifier block is removed and the output of the current last convolutional layer is used to create a descriptor. We propose to use this procedure for the Xception [2] model. Compared to GoogLeNet (InceptionV1), Xception uses improved Inception modules. This leads to an improved performance on the ImageNet [14] challenge. We refer to this new network as Xception Conv-Net. This Xception Conv-Net has $K = 2048$ filters in its last layer, thus yielding 2048-dimensional fine descriptors. Furthermore, with an input size of 512×256 these filters have a dimensionality of 16×8.

Third, we propose to improve the coarse descriptors. These are key in the C2F retrieval framework, as they enable the fast querying of the gallery. A further improvement can be accomplished by dimension reduction. We propose a dimensionality of 8. In this manner the average query time is reduced through a lower number of comparisons. Furthermore, as the PCA removes noise from the descriptors, an increase in quality is accomplished. This results in an enhanced accuracy and system performance.

4 Experimental setup

Models are trained on the train subsets of the Market-1501 [23] and CUHK03 (detected) [10] datasets, results are reported on the respective test sets.

Market-1501 [23] consists of 32,668 images with 1501 labeled identities. These are split into a train set containing 12,936 images with 751 identities and a test set containing 19,732 with 750 identities. Additionally 500,000 gallery distractor images are available to construct an extended test set with a size of 520k.

CUHK03 (detected) [10] contains 13,164 automatically detected images of 1360 pedestrians. These are subdivided into train and test subsets. Previously, a common approach was to select 100 identities for the test set and using the remaining identities for the train set. However, we use the more realistic and significantly more difficult approach proposed by Zhong et al [25]. With this approach the detected subset contains 7365 training images, 1400 query images and 5332 gallery images.

As evaluation metric we report the Cumulative Matching Characteristic (CMC) Rank-r and the mean of average precisions (mAP). The Rank-r indicates the probability that at least one matching image is in the first r positions of the ranked gallery. We mainly use the Rank-1. The mAP is calculated by taking the mean value of the average precisions for all of the performed queries.

The models in the experiments are initialized with ImageNet pre-trained weights without freezing any layers. For such networks it is a common practice to normalize images by subtracting the standard values $104, 117, 123$ from the red, blue and green channels respectively. Furthermore, data augmentation is done by flipping the images in training data sets with a chance of 50%.

5 Experiments and results

For the experiments, a machine was used with an Nvidia Titan X GPU and an Intel Core i7-4790 CPU. All models were implemented in Keras [1].

5.1 Automatic optimizers

In order to increase the accuracy of the net as well as allowing automated adjustment of the learning rate during the training, we propose to use the Adadelta optimizer. To demonstrate the effectiveness of Adadelta the ResNet50 [6] model is trained for 50 epochs on Market-1501 using various optimizers with their default settings [1]. Images are resized to 224×224, the standard image size for ImageNet. These are fed to the network in batches of 16. This batch size was experimentally shown to yield a higher accuracy than batches with 8 images. However, our system did not permit a larger batch size than 16. Note that Conv-Net and/or the C2F framework is not used. Results are presented in Table 1.

From our experiment in which 7 optimizers are tested Adadelta achieves the highest accuracy, both scores are with an absolute difference of 12.7% (mAP) and 8.7% (Rank-1) considerably higher than those of runner up AdaMax. Moreover, Adadelta exceeds the 51.48% mAP of ResNet50 trained using SGD as presented by Yao et al. [21], which is trained with better than default settings.

Table 1: Performance of ResNet50, trained with various optimizers on Market-1501

Optimizer	mAP	Rank-1	Rank-20	Optimizer	mAP	Rank-1	Rank-20
RMSprop [20]	29.8%	58.5%	90.0%	SGD [13]	46.4%	70.5%	93.9%
Adam [9]	31.3%	57.1%	87.6%	AdaMax [9]	46.5%	72.0%	94.1%
Nadam [3]	36.1%	63.6%	90.9%	Adadelta [22]	59.2%	80.7%	96.6%
Adagrad [4]	36.7%	62.0%	91.5%				

5.2 Base-models for Conv-Net

Constructing Conv-Net+C2F using Xception instead of GoogLeNet improves its accuracy. The coarse dimensionality of the descriptors is reduced with PCA to 128 as suggested by Yao et al [21]. Training is performed for 75 epochs on Market-1501 using a batch size of 16 and Adadelta as optimizer.

The input image size determines the dimensionality of the final output filters. An image size of 512×256 is used as much as possible for training these Conv-Net variants; this was suggested by Yao et al. This results in a final Conv-Net output of 16×8 for most models. This ensures that the models are tested on the same image sizes and at the same time supports in the construction of the descriptors. Especially the construction of the fine descriptor depends on the fact that the number of rows of the final output filters is evenly divisible by 4.

Unfortunately, this approach is not suitable for InceptionV3 and Inception-ResNetV2, since these setting would yield a final output of 14×6. To alleviate this problem, the image size is set to 586×299 for these two networks, which does result in a final output of 16×8 while respecting the $2:1$ height-width ratio as much as possible. Note that changing this size either violates the aspect ratio even more or results in a different sized final output. Moreover, MobileNet [8] requires its inputs to be square and no larger than 224×224. Results are presented in Table 2. Using Xception as a basis-network for constructing Conv-Net ('Xception Conv-Net') shows improved performance.

Table 2: Performance of various models as a basis for Conv-Net+C2F

Base-model	mAP	Rank-1	Base-model	mAP	Rank-1
GoogLeNet, Yao [21]	64.6%	84.6%	InceptionV3 [19]	61.4%	82.1%
VGG19 [15]	32.0%	53.5%	InceptionResNetV2 [18]	63.3%	84.4%
VGG16 [15]	34.0%	55.3%	ResNet50 [6]	64.8%	86.4%
MobileNet [8]	54.4%	81.1%	Xception [2]	66.5%	85.8%

5.3 Coarse descriptor dimensionality reduction

Using PCA to reduce the dimensionality of the coarse descriptors improves the accuracy of Xception Conv-Net+C2F. To demonstrate this, the performance of the model on Market-1501 and CUHK03 (detected) is measured with a coarse descriptor dimensionality range of 2^x with $1 \leq x \leq 8$. Training on the respective datasets is performed using Adadelta with a batch size of 16 for 75 epochs. Results are listed in Table 3.

The results clearly illustrate that 8-dimensional coarse descriptors offer the best mAP performance on Market-1501 and 4-dimensional coarse descriptors perform best on the smaller CUHK03 (cf. Table 3). The former settings are recommended for large gallery sizes where a high mAP is of importance. Furthermore, the mAP/Rank-1 scores with 8 (or 4) dimensions outperforms the

state-of-the-art method of Sun et al. [16] (37.83%/41.50%). This indicates that our approach generalizes to multiple datasets. Note that the new training/testing protocol for CUHK03 is used here. Using the old protocol Sun et al. obtained a mAP of 81.8% and Rank-1 of 84.8%. These accuracies are state-of-the-art as they outdo most other published works [16]. Our method outperforms theirs on the CUHK03 with the new protocol and by the principle of transitivity it is therefore acceptable to assume that our method would achieve state-of-the-art on the old protocol as well.

Table 3: Effect of coarse PCA descriptor dimensionality for Xception Conv-Net+C2F

	Market-1501		CUHK03	
dim	mAP	Rank-1	mAP	Rank-1
256	67.7%	86.6%	41.3%	41.9%
128	67.7%	86.6%	41.3%	41.9%
64	67.8%	86.6%	41.3%	41.9%
32	68.3%	86.6%	41.5%	41.9%
16	69.7%	86.5%	42.0%	41.8%
8	72.4%	85.7%	42.6%	41.9%
4	72.2%	81.6%	43.0%	41.6%
2	61.8%	64.6%	40.9%	38.4%

5.4 Market-1501 extended test data

Xception Conv-Net+C2F obtains both a high accuracy and a fast retrieval speed on a dataset with a large gallery containing many imperfections such as misaligned bounding-boxes and false-positives. The in Sec. 5.2 trained Xception Conv-Net is applied to the 520k Market-1501 images, results are listed in Table 4. Xception Conv-Net with 8-dimensional coarse descriptors improves on the mAP of Yao et al. [21] and also reduces the average query time by half. Furthermore, compared to only using GoogLeNet it is a factor 12 times faster.

Table 4: Performance of Xception Conv-Net+C2F (Our method) on extended Market

model	coarse dim	mAP	Rank-1	query time(ms)
GoogLeNet, Yao [21]	N/A	36.38%	56.05%	960
Conv-Net +C2F, Yao [21]	128	46.74%	64.58%	180
Xception Conv-Net+C2F (Ours)	128	58.0%	79.4%	152
Xception Conv-Net+C2F (Ours)	32	61.7%	79.2%	102
Xception Conv-Net+C2F (Ours)	8	63.8%	67.9%	81

5.5 Xception Conv-Net+C2F component-wise contribution

What is the influence of Xception Conv-Net and coarse-to-fine search on the obtained accuracy? And how does this compare to the state-of-the-art? To test this three different configurations are used. A baseline Xception network is compared to the Xception Conv-Net variant without C2F, with C2F using 128 coarse descriptors and another using 8-dimensional coarse descriptors.

The baseline Xception model is trained on the Market-1501 dataset for 75 epochs using Adadelta. Its descriptors are created by applying GAP to the 2048 filters of the last convolutional layer during testing, consequently these are 2048-dimensional. Xception Conv-Net is used in similar fashion as is done in Section 5.2 for the other models. For the variant without the C2F framework, only the coarse descriptors are used and PCA is not applied. Results for Conv-Net+C2F, with coarse descriptors having 128 or 8 dimensions, are taken from Section 5.3. Results are presented in Table 5.

The results show that all components of Xception Conv-Net+C2F improve its performance. The use of the Xception Conv-Net design instead of plain Xception contributes the most, with an absolute increase in mAP of 6.7% and in Rank-1 of 4.8%. Moreover, the use of the C2F framework increases the performance compared to using just Xception Conv-Net. This amounts to a further increase in mAP of 3.5% and in Rank-1 of 2.8%. Finally, using 8-dimensional coarse descriptors gives the best performance, this amounts to an absolute increase of the mAP with 4.7%. These results indicate that all components are necessary to achieve the reported state-of-the-art accuracy on Market-1501 (Table 5).

Table 5: Xception Conv-Net+C2F (ours) component-wise single-query results on Market-1501 compared to the state-of-the-art

Model (coarse dim)	mAP	Rank-1	Model (coarse dim)	mAP	Rank-1
Sun [16]	62.1%	82.3%	Zheng [24]	66.07%	83.97%
Zhong [25]	63.63%	77.11%	Hermans [7]	69.14%	84.92%
Yao [21]	64.6%	84.6%	Xception	57.5%	79.0%
Lin [12]	64.67%	84.29%	Xception Conv-Net	64.2%	83.8%
Li [11]	65.5%	85.1%	Xception Conv-Net +C2F (128)	67.7%	86.6%
Geng [5]	65.60%	83.75%	Xception Conv-Net +C2F (8)	72.4%	85.7%

6 Conclusion

For fast and accurate person re-id we propose to train Xception Conv-Net+C2F using Adadelta as optimizer with a batch size of at least 16. Furthermore, it is recommended to create coarse descriptors with a dimensionality of 8. Results

on Market-1501 (single-query 72.4% mAP), its extended version (single-query 63.8% mAP) and CUHK03 with the new protocol (detected 42.6% mAP) show that this results in higher accuracy, lower retrieval time (only 81ms for 520k images) and faster experimentation (no manual tuning of optimizer needed).

References

1. Chollet, F., et al.: Keras (2017), https://github.com/fchollet/keras
2. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. arXiv:1610.02357 (2016)
3. Dozat, T.: Incorporating Nesterov momentum into Adam. openreview (2016)
4. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. JMLR **12**(Jul), 2121–2159 (2011)
5. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person re-identification. arXiv:1611.05244 (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR. pp. 770–778 (2016)
7. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv:1703.07737 (2017)
8. Howard, A., Zhu, M., Chen, B., et al.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861 (2017)
9. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv (2014)
10. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: IEEE CVPR. pp. 152–159 (2014)
11. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. arXiv:1705.04724 (2017)
12. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. arXiv:1703.07220 (2017)
13. Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics pp. 400–407 (1951)
14. Russakovsky, O., Deng, J., Su, H., et al.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3), 211–252 (2015)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
16. Sun, Y., et al.: SVDNet for pedestrian retrieval. arXiv (2017)
17. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE CVPR (2015)
18. Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arxiv:1602.07261
19. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE CVPR. pp. 2818–2826 (2016)
20. Tieleman, T., Hinton, G.: RMSprop Gradient Optimization. Neural Networks for Machine Learning (2015), http://www.cs.toronto.edu/
21. Yao, H., et al.: Large-scale person re-identification as retrieval. ICME (2017)
22. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv:1212.5701 (2012)
23. Zheng, L., Shen, L., Tian, L., et al.: Scalable person re-identification: A benchmark. In: IEEE ICCV. pp. 1116–1124 (2015)
24. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. ICCV (2017)
25. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. arXiv:1701.08398 (2017)