Lightweight Residual Network for Person Re-identification

Reza Fuad Rachmadi, Supeno Mardi Susiki Nugroho, and I Ketut Eddy Purnama
Dept. of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia 60111
E-mail: {fuad,mardi}@its.ac.id, ketut@te.its.ac.id

Abstract. In this paper, we investigated several lightweight convolutional neural network (CNN) classifiers for person re-identification problems. Person re-identification can be described as finding a person by performing a query with a specific person image on some defined gallery of people images. We construct the lightweight CNN classifier by utilizing the lightweight residual network architectures originally used for the CIFAR dataset. The goal is to design a lightweight CNN classifier with a maximum of 3 million number of parameters that produce good accuracy on person re-identification problems. The last layer of the residual network is detached and changed by two fully-connected layers, one for learning the discriminant features, and the other fully-connected layer for calculated the final classification score in the training process. Experiment results show that the ensemble of lightweight residual networks achieved a good performance on Market-1501 (rank-1 accuracy of 89.46% with mAP of 84.48% for single-query and rank-1 accuracy of 92.23% with mAP of 88.07% for multi-query). Although the lightweight residual network does not achieve state-of-the-art performance, our analysis shows that the lightweight residual network has higher information density comparing with other state-of-the-art models. More information density means that the classifier is more efficient in terms of performance with the number of parameters in the classifier. Implementation of these experiments are available at https://github.com/rezafuad/person-reid-lightweight-residualnet.

1. Introduction
The development of a smart security system is one of the active research areas that accelerate very fast after the introduction of deep learning technology. One of the most implemented technologies of security system is a network of CCTV cameras placed in several areas of the city or public places for monitoring purposes. Searching specific objects, such as a person or car, in the CCTV video data is one of the tasks for a smart security system. Person re-identification problems can be described as finding a specific person on the entire people’s data extracted from the CCTV video in the same area and timestamp. In the last five years, the development of person re-identification models is in a very active state with several datasets [1, 2, 3, 4] and models [5, 6, 7, 8] had been proposed by the researchers. Person re-identification problems are equivalent to the image retrieval problem that specific for person images. Given a query of a specific image person, the person re-identification model returns all of the images in the gallery that match the query image. Although the development of the person re-identification models has been produced several state-of-the-art models, the investigating of lightweight CNN
models for person re-identification problems are not yet addressed. A lightweight CNN classifier for person re-identification problems is useful for deploying the models in the edge computing scheme which have limited hardware specification choices.

In this paper, we investigated several lightweight residual CNN architecture as a solution to person re-identification problems. We constructed the lightweight residual CNN classifier based on the residual network [9] that used to solve CIFAR image classification problems. We removing the fully-connected layers of the original classifier and added two new fully-connected layers with different configuration from the original classifier. The contributions we make can be listed as follows.

- We investigated several lightweight residual network classifiers for person reidentification problems. The classifier constructed using residual network architecture that was originally used for the CIFAR dataset. Five different residual network architectures are used for the experiments, including ResNet20, ResNet32, ResNet44, ResNet56, and ResNet110.

- From the experiments, we proved that lightweight residual network classifiers with ensemble configuration can also produce very good performance comparing with other state-of-the-art methods on the Market-1501 dataset. Although the lightweight is not achieved state-of-the-art performance, the information density shows that the lightweight residual network has more information density comparing with other state-of-the-art models which proved that the classifier is more efficient in term of performance with the number of parameters in the classifier.

- We perform an ablation study for several lightweight residual networks to see the performance impact of the classifier if some of the features are removed or added to the classifier. Experiments show that reducing the number of neurons on the fully-connected layer can reduce the overfitting problem and improve the performance of the classifier.

The rest of the paper is organized as follows. The setting of the experiments of lightweight CNN classifiers is discussed in section two. The ablation study on the Market-1501 dataset is discussed in section three. Additional experiments by adjusting several parameters/method, including dropout rate, re-ranking, and multi-scale, is discussed in section four. Finally, we conclude the experiments in section five.

2. Experiments Setup

In this section, we describe the experiment setting, including the classifier we use for the experiments, the dataset used in the experiments, the training, and the testing process. All of the experiments are done using PyTorch framework and NVIDIA RTX 2080 TI with 11 GB memory.
2.1. Lightweight CNN

Figure 1 shows the simple diagram of our proposed lightweight CNN classifier. We construct the lightweight CNN classifier based on residual network architecture that used to solve CIFAR image classification problems. The final fully-connected of the residual network is removed and two new fully-connected layers is added at the end of the classifier. The first fully-connected is used to learn discriminant features while the last fully-connected layer is used to classify the identity of the person in the training data. To ensure that the first fully-connected layer learns good discriminant features, no activation function is attached after the first fully-connected layer. In the testing phase, the last final fully-connected layer is removed and the features used for the retrieval process are extracted from the first fully-connected layer. Unlike the original residual network, our lightweight CNN classifier use an input resolution of $32 \times 64$ instead of $32 \times 32 \times 3$. We use five different residual network, including ResNet20, ResNet32, ResNet44, ResNet56, and ResNet110. Although the number of layers is very deep (e.g. 110 layers), the number of parameters of the deepest classifier is still under 3 million parameters.

Is there any other lightweight CNN architectures that can be used for solving the person re-identification problem? Based on our study, some other lightweight CNN architectures can be used for solving the person re-identification problems but in this paper, we focus the experiments for residual-based lightweight CNN architecture.

2.2. Person Re-identification Dataset

To evaluating the performance of the lightweight classifier, we mainly use the Market-1501 dataset which consists of 1,501 unique identities captured using six different cameras. The dataset was divided into 750 identities for the training process and 751 identities for the testing process with a total number of 32,668 bounding boxes. Each identity is unique and no overlapping between the training and testing process. Each person image is cropped and normalize to $128 \times 64$ using a combination of DPM (Deformable Parts Model) and handcrafted method. Figure 2 shows several samples of person images in Market-1501 person re-identification dataset.

![Figure 2. Some person images from Market-1501 person re-identification datasets.](image-url)
In this paper, we use a different resolution for our classifier input. We use resolution of 64×32 for the training and testing process which half of the original Market-1501 resolution. The training dataset of Market-1501 is reorganized such that the data can be used for the training process using classification loss function with one person image per identity is used for the validation process. We choose the classifier with the highest accuracy based on the validation set for further evaluation.

2.3. Training and Testing Process
To ensure the performance analysis is reliable, we perform the training and testing for five iterations and averaging the evaluation metric. Three evaluation metric is used for evaluating the performance of the classifier, including rank, mAP, and information density.

The training process is done for 100 epochs with the learning rate initialize at 0.1 and decreased by a factor of 0.1 when reaching 40 and 80 epoch. All of the convolutional layer weights were initialize using weights from the same CNN architecture trained using CIFAR datasets [10]. The different learning rate is used for the convolutional layers with an initial value of 0.01 with the same policy as used in the global learning rate. We use two different data augmentation process, random crop and random erasing data augmentation follow by subtracting using the mean of the training data. Figure 3 shows some examples of data augmentation images using those two methods. In the first fully-connected layer we added a dropout layer with several different dropout ratios. The cross-entropy loss function is used to train the classifier with softmax activation at the end of the classifier.

In the testing process, the features for the retrieval process are taken from the first fully-connected layer. Same as in the training process, the input image is resized to 64×32 and subtracted by the mean of the training data. The similarity score is computed by correlating two extracted features and sorted the results from low to high. A low score means that the two features are correlated and vice-versa.

3. Ablation Study
To evaluate the ResNet-CIFAR lightweight classifier for person re-identification problems, we perform an ablation study using three different protocols, by changing the random erasing
Figure 4. Results from the ablation study of the lightweight CNN classifier for person re-identification problem using Market-1501, (a) random erasing portion, (b) size of FC (or size of features), and (c) impact of input resolution.

portion, by changing the size of the first fully-connected layer, and by changing the input resolution of the classifier. We use mAP (mean average precision) as the main metric in our ablation study. Figure 4 shows the ablation study results in graphic form.

In the first ablation study, we change the random erasing portion from 0 to 50% with an increment of 10%. As shown in Figure 4-(a), the random erasing data augmentation proved can improve the classifier with appropriate portion value. After the random erasing portion of 30%, the performance is not consistent for all classifiers due to the large erasing portion of the input image which effected the classification performance.

Several different sizes of the first fully-connected layer (FC-size) in the classifier is our next ablation study. In this and next experiments, we only use ResNet56 and ResNet110 which in the last experiments achieves first and second highest mAP. Figure 4-(b) shows the plot of FC-size with the performance of the classifier (in mAP). As shown in Figure 4-(b), the best mAP is achieved using an FC-size of 128 which consider the lowest FC-size configuration in our experiments. After analyzing the training and validation process, we concluded that bigger FC-size lead to overfitting problems which effected the testing performance of the classifier.
In the last ablation study, we try several different input resolutions for the classifier and measure the mAP metric. Figure 4-(c) shows the plot of FC-size with the performance of the classifier (in mAP) using three different input resolutions. As shown in Figure 4-(c), input resolutions of 128×64 and 64×32 have a similar performance with a very narrow gap. The opposite results happen when the input resolution reduces to 32×16 which produces around 50% drop of mAP comparing with other input resolutions.

4. Results
The Market-1501 dataset testing process is split into two different protocols, single-query testing protocol, and multi-query testing protocol. The single-query testing protocol done by given only one query image to the classifier while the multi-query is given more than one query image. Naturally, the multi-query results will be higher than the single query due to the more query samples expose to the classifier.

Based on the results in our ablation study, we decided to use an input resolution of 64×32, a droprate of 0.5, a random erasing portion of 0.3, and an FC-size of 128. We only choose ResNet56 and ResNet110 CNN architecture for further experiments. The hyperparameter used in the training process is also the same as in the ablation study. Additional re-ranking method [11] and SPP (Spatial Pyramid Pooling) [12] is used in the experiments which proved can improve retrieval accuracy (especially rank-1 and mAP). Table 1 shows the summary of our experiments using ResNet56 and ResNet110 with additional RE (Random Erasing) and re-ranking method.

As shown in Table 1, the additional methods (SPP and Re-ranking) can increase the performance of the classifier by around 1% for SPP and around 3% for reranking. The best performance is achieved using a combination of ResNet110, random erasing, SPP, and re-ranking method. The different results for rank-5 accuracy in which the best performance is not achieved by the classifier with all additional methods. Based on Table 1 results, the phenomena also appear for other classifiers when the testing processes using re-ranking method.

4.1. Ensemble
To increase the performance of the classifier, we also conducted testing process using ensemble of ResNet56 and ResNet110 classifiers. Ensemble of two or more CNN classifiers proved can increase the performance of the classifier by around 2%. We use a simple average ensemble
Table 2. Experiment using an ensemble of ResNet56 and ResNet110 on Market-1501 datasets (averaging from five iterations).

| Model                  | Single Query (%) | Multi Query (%) |
|------------------------|------------------|-----------------|
|                        | rank-1 | rank-5 | mAP  | rank-1 | rank-5 | mAP  |
| Ensemble-RE            | 86.95±0.61 | 94.99±0.20 | 69.94±0.46 | 91.52±0.37 | 96.97±0.14 | 78.01±0.42 |
| Ensemble-SPP-RE        | 86.97±0.49 | 94.98±0.18 | 69.87±0.19 | 91.16±0.38 | 96.79±0.18 | 77.60±0.26 |
| Ensemble-RE-Rerank     | 89.23±0.28 | 93.96±0.44 | 84.45±0.75 | **92.39±0.27** | 96.09±0.31 | **88.24±0.38** |
| Ensemble-SPP-RE-Rerank | **89.46±0.44** | 94.08±0.47 | **84.48±0.24** | 92.23±0.37 | 96.01±0.27 | 88.07±0.07 |

Table 3. Comparison of our lightweight classifier with several state-of-the-art models.

| Years | Method                     | #Params (million) | Single (%) | Market-1501 | Multi (%) | Inf. Den. |
|-------|----------------------------|-------------------|------------|-------------|-----------|-----------|
|       |                            |                   | r = 1 mAP  | Inf. Den.   | r = 1 mAP | Inf. Den. |
| 2015  | BoW + KISSME [2]           | n/a               | 44.42      | 20.76       | -         | -         | -         |
| 2016  | SL [13]                    | n/a               | 51.90      | 26.35       | -         | -         | -         |
| 2016  | DNS [14]                   | n/a               | 61.02      | 35.68       | 71.56     | 46.03     | -         |
| 2016  | SSDAL [15]                 | ~61               | 39.4       | 19.6        | 0.321     | 49.0      | 25.8      | 0.422     |
| 2017  | CAN [16]                   | ~50               | 60.3       | 35.9        | 0.718     | 72.1      | 47.9      | 0.958     |
| 2017  | Fisher Net [17]            | 5.2               | 48.15      | 29.94       | 5.757     | -         | -         | -         |
| 2017  | Verif-Iden [18]            | ~23               | 79.51      | 59.87       | 2.603     | 85.84     | 70.33     | 3.057     |
| 2018  | PAN [19]                   | ~45               | 88.57      | 81.53       | 1.811     | 91.45     | 87.44     | 1.943     |
| 2019  | Spatial-Temporal [20]      | ~30               | 98.0       | 95.5        | 3.183     | -         | -         | -         |
| 2019  | PyrNet [21]                | 35.9              | 94.6       | 91.4        | 2.545     | 96.1      | 94.0      | 2.618     |
| 2019  | ARP [22]                   | ~23               | 87.04      | 66.89       | 2.908     | -         | -         | -         |
| 2020  | VA-ReID [23]               | 52.64             | 96.79      | 95.43       | 1.812     | -         | -         | -         |
| 2020  | SARL [24]                  | ~32               | 96.1       | 88.0        | 2.750     | -         | -         | -         |
| 2020  | RGA-SC [25]                | 56.4              | 96.1       | 88.4        | 1.567     | -         | -         | -         |
|       | Our (Single)               | 1.8               | 87.67      | 82.04       | 45.57     | 91.48     | 86.27     | 47.93     |
|       | Our (Ensemble)             | 2.7               | 89.46      | 84.48       | 31.29     | 92.23     | 88.07     | 32.62     |

Method between those two models. Table 2 shows the results of ensemble classifiers using the Market-1501 person re-identification dataset. In single-query testing, the best performance is achieved using Ensemble of ResNet56 and ResNet110 with additional SPP method, random erasing, and reranking. In multi-query testing, the best performance is achieved using Ensemble of ResNet56 and ResNet110 with additional random erasing and reranking. The same phenomena appear in previous experiments is also appear in this experiment in which the ensemble classifier with an additional reranking method produces a lower rank-5 accuracy comparing with the ensemble classifier that does not use the re-ranking method.
4.2. Comparison

We added information density metrics which measure the effectiveness of the classifier in term of performance (in mAP) and the number of parameters in the classifier. Let $f_{m\text{AP}}$ is measured performance of the classifier (in mAP) and $f_p$ is number of parameters (in million) in the classifier, the information density can be calculated as follows.

$$I_{\text{den}} = \frac{f_{m\text{AP}}}{f_p}$$

(1)

Table 3 shows the performance of our lightweight residual network classifier comparing with several other methods. As shown in Table 3, although our lightweight residual network classifier is not achieved a state-of-the-art performance but after we compute the information density, our lightweight residual residual network classifier have very high with around 10 to 20 times higher than other methods. In terms of single query rank-1 and mAP metric, the lightweight residual network classifier is around 10% lower than the state-of-the-art method while in terms of multi-query rank-1 and mAP metrics, the lightweight residual network classifier is 4-8% lower than the state-of-the-art classifier.

5. Conclusion

We have investigated several lightweight CNN architectures for person re-identification problems. The lightweight CNN architecture formed using ResNet-CIFAR CNN architecture by removing the last fully-connected layer and adding two fully-connected layers. The first fully-connected layer is used for learning the discriminant features and the second fully-connected layer for the final classification score in the training process. To avoid dead neurons in the first fully-connected, no activation function is added in the first fully-connected layer. In the testing process, we removed the second fully-connected layer and extracted the features for the retrieval process from the first fully-connected layer. Based on the experiments using Market-1501 person re-identification, the best performance achieved using Ensemble of ResNet56 and ResNet110 with rank-1 accuracy of 89.46%, mAP of 84.48% in single query testing and rank-1 accuracy of 92.23%, mAP of 88.07% in multi-query testing. Although the lightweight ResNet-CIFAR CNN classifier not achieved state-of-the-art performance, the information density of the lightweight ResNet-CIFAR CNN classifier is very high which proved that the classifier is more efficient comparing with other methods.

Deeper investigations of the lightweight CNN classifier on other person re-identification datasets is one of our concerns for future work. A combination of the lightweight CNN classifier with deep metric loss also worthed to investigate because the deep metric learning, such as sphereface or center-loss, proved can increase the performance of the classifier on face recognition problems.

Acknowledgments

This research is partially supported by Indonesia Ministry of Education and Culture Research Grant (under PDUPT Scheme) No. 1206/PKS/ITS/2020.

References

[1] Li W, Zhao R, Xiao T and Wang X 2014 Deepreid: deep filter pairing neural network for person re-identification Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 152–9
[2] Zheng L, Shen L, Tian L, Wang S, Wang J and Tian Q 2015 Scalable person re-identification: a benchmark Proc. of the IEEE Int. Conf. on Computer Vision pp 1116–24
[3] Ristani E, Solera F, Zou R, Cucchiara R and Tomasi C 2016 Performance measures and a data set for multi-target, multi-camera tracking European Conf. on Computer Vision workshop on Benchmarking Multi-Target Tracking pp 17–35
[4] Zheng Z, Zheng L and Yang Y 2017 Unlabeled samples generated by gan improve the person re-identification baseline in vitro Proc. of the IEEE Int. Conf. on Computer Vision pp 3774–82
[5] Sun Y, Zheng L, Yang Y, Tian Q and Wang S 2018 Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline) Proc. of the European Conf. on Computer Vision (ECCV) pp 480–96
[6] Zheng Z, Yang X, Yu Z, Zheng L, Yang Y and Kautz J 2019 Joint discriminative and generative learning for person re-identification Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 2138–47
[7] Fu Y, Wei Y, Wang G, Zhou Y, Shi H and Huang T S 2019 Self-similarity grouping: a simple unsupervised cross domain adaptation approach for person re-identification Proc. of the IEEE Int. Conf. on Computer Vision pp 6112–21
[8] Zhang X, Cao J, Shen C and You M 2019 Self-training with progressive augmentation for unsupervised cross-domain person re-identification Proc. of the IEEE Int. Conf. on Computer Vision pp 8222–31
[9] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 770–8
[10] Idelbayev Y Proper ResNet Implementation for CIFAR10/CIFAR100 in PyTorch retrieved on 31 March 2020 from https://github.com/akamaster/pytorch
[11] Yu R, Zhou Z, Bai S and Bai X 2017 Divide and fuse: A re-ranking approach for person re-identification Proc. of the British Machine Vision Conf. (BMVC) ed Kim T K, Zafeiriou S, Brostow G and Mikolajczyk K (BMVA Press) pp 135.1–13 ISBN 1-901725-60-X dOI: 10.5244/C.31.135
[12] He K, Zhang X, Ren S and Sun J 2015 Spatial pyramid pooling in deep convolutional networks for visual recognition IEEE Transactions on Pattern Analysis and Machine Intelligence 37 pp 1904–16
[13] Chen D, Yuan Z, Chen B and Zheng N 2016 Similarity learning with spatial constraints for person re-identification Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 1268–77
[14] Zhang L, Xiang T and Gong S 2016 Learning a discriminative null space for person re-identification Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 1239–48
[15] Su C, Zhang S, Xing J, Gao W and Tian Q 2016 Deep attributes driven multi-camera person re-identification European Conf. on Computer Vision (New York: Springer) pp 475–91
[16] Liu H, Feng J, Qi M, Jiang J and Yan S 2017 End-to-end comparative attention networks for person re-identification IEEE Transactions on Image Processing 26 pp 3492–506
[17] Wu L, Shen C and Van Den Hengel A 2017 Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification Pattern Recognition 65 pp 238–50
[18] Zheng Z, Zheng L and Yang Y 2017 A discriminatively learned cnn embedding for person reidentification ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14 pp 1–20
[19] Zheng Z, Zheng L and Yang Y 2018 Pedestrian alignment network for large-scale person re-identification IEEE Transactions on Circuits and Systems for Video Technology 29 pp 3037–45
[20] Wang G, Lai J, Huang P and Xie X 2019 Spatial-temporal person re-identification Proc. of the AAAI Conf. on Artificial Intelligence vol 33 pp 8933–40
[21] Martinel N, Luca Foresti G and Micheloni C 2019 Aggregating deep pyramidal representations for person re-identification Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops pp 0–0
[22] Lin Y, Zheng L, Zheng Z, Wu Y, Hu Z, Yan C and Yang Y 2019 Improving person re-identification by attribute and identity learning Pattern Recognition 95 pp 151–61
[23] Zhu Z, Jiang X, Zheng F, Guo X, Huang F, Sun X and Zheng W 2020 Viewpoint-aware loss with angular regularization for person re-identification. AAAI pp 13114–21
[24] Jin X, Lan C, Zeng W, Wei G and Chen Z 2020 Semantics-aligned representation learning for person re-identification. AAAI pp 1177–80
[25] Zhang Z, Lan C, Zeng W, Jin X and Chen Z 2020 Relation-aware global attention for person re-identification Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition pp 3186–95