


RESEARCH

Open Access



Exploiting prunability for person re-identification

Hugo Masson^{1†}, Amran Bhuiyan^{1,2*†} , Le Thanh Nguyen-Meidine^{1†}, Mehrsan Javan², Parthipan Siva², Ismail Ben Ayed¹ and Eric Granger¹

*Correspondence:

amran.apece@gmail.com

[†]Hugo Masson, Amran Bhuiyan and Le Thanh Nguyen-Meidine contributed equally to this work.

¹LIVIA, Department of Systems Engineering, École de technologie supérieure, Montreal, Canada

²SLIQ Labs, Sportlogiq Inc., Montreal, Canada

Abstract

Recent years have witnessed a substantial increase in the deep learning (DL) architectures proposed for visual recognition tasks like person re-identification, where individuals must be recognized over multiple distributed cameras. Although these architectures have greatly improved the state-of-the-art accuracy, the computational complexity of the convolutional neural networks (CNNs) commonly used for feature extraction remains an issue, hindering their deployment on platforms with limited resources, or in applications with real-time constraints. There is an obvious advantage to accelerating and compressing DL models without significantly decreasing their accuracy. However, the source (pruning) domain differs from operational (target) domains, and the domain shift between image data captured with different non-overlapping camera viewpoints leads to lower recognition accuracy. In this paper, we investigate the prunability of these architectures under different design scenarios. This paper first revisits pruning techniques that are suitable for reducing the computational complexity of deep CNN networks applied to person re-identification. Then, these techniques are analyzed according to their pruning criteria and strategy and according to different scenarios for exploiting pruning methods to fine-tuning networks to target domains. Experimental results obtained using DL models with ResNet feature extractors, and multiple benchmarks re-identification datasets, indicate that pruning can considerably reduce network complexity while maintaining a high level of accuracy. In scenarios where pruning is performed with large pretraining or fine-tuning datasets, the number of FLOPS required by ResNet architectures is reduced by half, while maintaining a comparable rank-1 accuracy (within 1% of the original model). Pruning while training a larger CNNs can also provide a significantly better performance than fine-tuning smaller ones.

Keywords: Deep learning, Convolutional neural networks, Complexity, Pruning, Domain adaptation, Person re-identification

1 Introduction

Deep learning (DL) architectures like the convolutional neural network (CNN) have achieved state-of-the-art accuracy across a wide range of visual recognition tasks, at the expense of growing complexity (deeper and wider networks) that require more training samples and computational resources. In order to deploy these architectures on

compact platforms with limited resources (e.g., embedded systems, mobile phones, portable devices), and for real-time processing (e.g., video surveillance and monitoring, virtual reality), their time and memory complexity and energy consumption should be reduced [1]. Consequently, there is a growing interest in effective methods able to accelerate and compress deep networks.

Providing a reasonable trade-off between accuracy and efficiency has become an important concern in person re-identification (ReID), a key function needed in a wide range of video analytics and surveillance applications. Systems for person ReID typically seek to recognize the same individuals that previously appeared over a non-overlapping network of video surveillance cameras (see illustration in Fig. 1). These systems face many challenges in real-world applications that are related to either the image data or the network architecture. Data-related challenges that affect the ReID accuracy include the limited availability of annotated training data, ambiguous annotations, domain shifts across camera viewpoints, limitations of person detection and tracking techniques, occlusions, variations in pose, scale and illumination, and low-resolution images.

To address these issues, state-of-the-art DL models (e.g., deep Siamese networks) for person ReID often rely CNNs for feature extraction to learn an embedding in end-to-end fashion, where similar image pairs (with the same identity) are close to each other, while dissimilar image pairs (with different identities) are distant from each other [2–10]. While state-of-the-art approaches can provide a high level of accuracy, achieving this performance comes with the cost of millions or even billions of parameters, a challenging training procedure, and the requirement for GPU acceleration. For instance, the ResNet50 CNN [11], with its 50 convolutional layers, contains about 23.5M parameters (stored in 85.94MB of memory) and requires 6.3 billion floating point operations (FLOPs) to process a color image of size $256 \times 128 \times 3$. The complexity of these networks limits their deployment in many real-time applications, or on resource-limited platforms. Consequently,

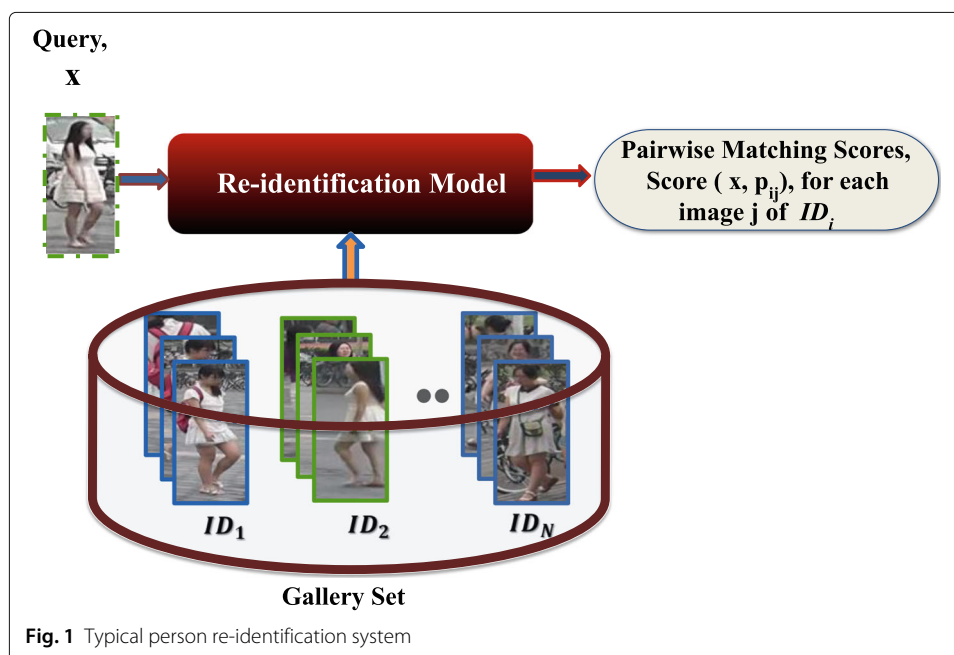


Fig. 1 Typical person re-identification system

there has been a great deal of interest by the computer vision and machine learning communities to develop methods able to accelerate and compress such networks, as well as other DL architectures, without compromising their predictive accuracy.

The time complexity of a CNN generally depends more on the convolutional layers, while the fully connected layers contain the most of the parameters (memory complexity). Therefore, the CNN acceleration methods typically target lowering the complexity of the convolutional layers, while the compression methods usually target reduced complexity of the fully connected layers [12, 13]. State-of-the-art approaches for acceleration and compression of deep neural networks can be divided into five categories—low-rank factorization, transferred convolutional channels, knowledge distillation, quantization, and pruning.

Low-rank factorization approaches [14–19] accelerate CNNs by performing matrix decomposition to estimate information parameters of a network. However, low-rank approaches suffer from a number of issues—computationally expensive matrix decomposition, layer-by-layer low-rank approximation that diminishes the possibility of global compression, and extensive model retraining to achieve convergence. Some network acceleration and compression approaches [20–23], categorized as transferred convolutional channels, design special structural convolutional channels to reduce the parameter space, which eventually improves computational efficiency, but transfer assumptions are sometimes too strong that makes the learning process unstable.

Knowledge distillation approaches [24–27] train a smaller or shallow deep network (the student) using distilled knowledge of a larger deep network, called teacher. These approaches can yield improvements in terms of sparsity and generalization of the student networks, but can only be applied to classification tasks and the bounded assumption of this approach leads to inferior performance while comparing to other types of approaches. A deep neural network can be accelerated by reducing the precision of its parameters. Using quantization approaches, each parameter of a network is represented with a reduced bit rate, either by reducing the precision, employing a lookup table, or combining similar values. Most of the quantization approaches [12, 28, 29] require extra computational time to access a look up table, or for decoding such that the original value is restored. In contrast, pruning seeks to reduce the number of connections or retrain either the whole or part of the network with a freshly trained replacement. Pruning methods typically focus on selecting and removing the weights or channels with the least impact on performance. Thus, in addition to accelerating and compressing the network, pruning methods can provide the additional benefits such as addressing the overfitting problem and thus improve generalization. Therefore, pruning approaches have drawn a great deal of attention from the network compression community. Challenges of pruning include the lack of data for pruning during the fine-tuning phase, the computational complexity associated with retraining after a pruning phase, and the reduction of capacity to learn of a model, which can impact the accuracy when the learning step is done on the pruned model.

This paper focuses on pruning techniques [13, 30–33] since they are among the most widely used for acceleration and compression of deep neural networks, and have been shown their effectiveness on well-known CNNs, and for several general image classification problems like CIFAR10, MNIST, and ImageNet. State-of-the-art pruning techniques can be categorized according to their pruning criteria to select channels, and to their

strategy to reduce channels, and are suitable for compressing DL models like Siamese networks for applications in person ReID. In particular, state-of-the-art techniques can be categorized using criteria based on weights and on feature maps. We also distinguish techniques according to pruning strategy; pruning techniques can also be distinguished among—those that (1) prune once and then fine-tune, (2) prune iteratively on trained model, (3) prune using regularization, (4) prune by minimizing the reconstruction errors, and (5) prune progressively.

This paper revisits the pruning techniques that are suitable for reducing the computational complexity of CNNs applied to person ReID. These techniques are then analyzed according to their pruning criteria and strategy and according to different design scenarios. Different design pipelines or scenarios are proposed to leverage these state-of-the-art pruning methods during pretraining and/or fine-tuning. A typical design scenario consists of four stages: (1) training a CNN with a large-scale dataset from the source domain (i.e., ImageNet), (2) prune the trained large model based on some criterion to select channels to be eliminated, (3) retrain the pruned network to regain the accuracy, and finally, (4) fine-tune the retrained network using a limited dataset from the target application. A common assumption with this design scenario is that training a large and over-parameterized CNNs, using a large-scale dataset, is necessary to provide a discriminant feature representation. The pruning process used to select and reduce the network will yield a set of redundant channels that does not significantly reduce accuracy. Under this scenario, a CNN for ReID would therefore over-train on a smaller network from scratch [33–36]. Thus, most of the approaches in literature tend to prune channels of a fine-tuned network, rather than a pretrained network. This paper presents other design scenarios that apply when pruning networks that have been pretrained on large dataset, and that require a fine-tuning to a given target domain.

Finally, this paper presents an extensive experimental comparison of different pruning techniques and relevant design scenarios on three benchmark person ReID datasets—the Market-1501 [37], CUHK03-NP [38], and DukeMTMC-reID [39]. Pruning techniques are compared in terms of accuracy and complexity on different DL architectures with ResNet feature extractors, with different ReID applications in mind.

The rest of the paper is organized as follows. Section 2 provides some background on DL models for person ReID. Section 3 provides a survey of the state-of-the-art techniques for pruning CNNs. Finally, Sections 5 and 6 described the experimental methodology (benchmark datasets, protocol, and performance measures) and comparative results, respectively.

2 Deep neural networks for person re-identification

State-of-the-art techniques for person ReID mostly rely on two types of losses: metric learning loss and multi-class classification loss. With the first type, a dataset with images from different individuals is learned using a Siamese network that optimizes a metric loss function (such as contrastive loss, triplet loss, quadruplet loss, hard-aware point-to-set (HAP2S) loss) [2, 4, 7, 40, 41] to provide a feature embedding for pairwise similarity matching. With the second type, ReID approaches based on multi-class classification loss (such as softmax or cross-entropy loss) [42–46] learn part-based local features to form more informative feature descriptor, also known as ID-loss in ReID community.

Table 1 Common loss functions applied in person ReID

Category	Loss function
Metric learning	Contrastive loss [4]: $\mathcal{L}_{CE} = \frac{1}{2N} \sum_{i=1}^N [(1 - y_i) d^2(\mathbf{f}_{1,i}, \mathbf{f}_{2,i}) + (y_i) \{\max(0, m - d^2(\mathbf{f}_{1,i}, \mathbf{f}_{2,i}))\}]$
	Triplet [47]: $\mathcal{L}_T = \frac{1}{N_T} \sum_{\substack{a,p,n \\ y_a=y_p \neq y_n}} [d(\mathbf{f}_a, \mathbf{f}_p) - d(\mathbf{f}_a, \mathbf{f}_n)]_+$
	Triplet loss with margin [7]: $\mathcal{L}_T = \frac{1}{N_T} \sum_{\substack{a,p,n \\ y_a=y_p \neq y_n}} [m + d(\mathbf{f}_a, \mathbf{f}_p) - d(\mathbf{f}_a, \mathbf{f}_n)]_+$
	Semi-hard triplet [41]: $\mathcal{L}_{TBH} = \frac{1}{N_s} \sum_{a=1}^{N_s} \left[m + \max_{y_p=y_a} d(\mathbf{f}_a, \mathbf{f}_p) - \min_{y_p \neq y_a} d(\mathbf{f}_a, \mathbf{f}_n) \right]_+$
	$\mathcal{L}_{quad} = \frac{1}{N} \sum_{\substack{a,p,n \\ y_a=y_p \neq y_n}} [m_1 + d(\mathbf{f}_a, \mathbf{f}_p) - d(\mathbf{f}_a, \mathbf{f}_n)]_+$
	Quadruplet [5]: $+ \frac{1}{N} \sum_{\substack{a,p,n,k \\ y_a=y_p \neq y_n \neq y_k}} [m_2 + d(\mathbf{f}_a, \mathbf{f}_p) - d(\mathbf{f}_n, \mathbf{f}_k)]_+$
HAP2S [48]: $\mathcal{L}_{HAP2S} = \frac{1}{N_s} \sum_{a=1}^{N_s} \left[m + \max_{y_p=y_a} d(\mathbf{f}_a, \mathbf{S}_p) - \min_{y_p \neq y_a} d(\mathbf{f}_a, \mathbf{S}_n) \right]_+$	
Magnet [49]: $\mathcal{L}_{mag} = -\frac{1}{N} \sum_{i=1}^N \left[\log \frac{e^{-\frac{1}{2\sigma^2} d(\mathbf{f}_i, \mu(\mathbf{f}_i)) - m}}{\sum_{k=1}^C e^{-\frac{1}{2\sigma^2} d(\mathbf{f}_i, \mu_k^i)}} \right]_+$	
Classification	Cross-entropy [2, 38, 50]: $\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\mathbf{W}_i^T \mathbf{f}_i}}{\sum_{k=1}^C e^{\mathbf{W}_k^T \mathbf{f}_i}}$
	Cosine Softmax [49]: $\mathcal{L}_{CCE} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\tilde{\mathbf{W}}_i^T \mathbf{f}_i}}{\sum_{k=1}^C e^{\tilde{\mathbf{W}}_k^T \mathbf{f}_i}}$
	Part-based cross-entropy [43–46]: $\mathcal{L}_{PCE} = \sum_{p=1}^P \mathcal{L}_{CE}^p$

Table 1 provides a summary of common loss functions from both categories applied in person ReID. For all the losses, d represents the Euclidean distance, m , m_1 , m_2 denotes the margin parameters, and $[\cdot]_+ = \max(\cdot, 0)$. In this table, $\mathbf{X} = \{\mathbf{x}\}_{i=1}^N$ is a training mini-batch with labels $\{y_i\}_{i=1}^N$. For contrastive loss, the network transforms the pair of input images $\mathbf{x}_{1,i}, \mathbf{x}_{2,i}$ into feature embeddings $\mathbf{f}_{1,i}, \mathbf{f}_{2,i}$. The labels are either $y_i = 0$ for positive pairs or $y_i = 1$ for negative pairs. For triplet loss, we sample $(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n)$ where the anchor and the positive \mathbf{x}_a are two images from the same person, while the negative \mathbf{x}_n is an image from another person. The corresponding feature embeddings are $(\mathbf{f}_a, \mathbf{f}_p, \mathbf{f}_n)$. For quadruplet loss, we sample $(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n, \mathbf{x}_k)$ where the anchor and positive \mathbf{x}_a are two images from the same person, while the negative and \mathbf{x}_k are the images from different persons. The corresponding feature embeddings are $(\mathbf{f}_a, \mathbf{f}_p, \mathbf{f}_n, \mathbf{f}_k)$. For HAP2S loss, \mathbf{S}_p and \mathbf{S}_n denote the set of positive and negative samples respectively with respect to the anchor \mathbf{f}_a ; for magnet loss, C is the number of classes (individual), μ is the sample mean of class y , and σ^2 is the variance of all samples away from their class mean. Belonging to classification loss category, in cross-entropy loss, \mathbf{W}_{y_i} is the weight vector of the fully connected layer with feature embedding \mathbf{f}_i . For cosine softmax loss, $\tilde{\mathbf{W}}_{y_i}$ and $\tilde{\mathbf{f}}_i$ denote the normalized weight and feature vector, respectively, and for part-based cross-entropy loss, \mathcal{L}_{CE}^p represents the cross-entropy loss of individual part, P .

2.1 Metric loss

The idea of using deep Siamese networks for biometric authentication and verification originates from Bromeley et al. [51], where two sub-networks with shared weights encode feature embeddings for pairwise matching between a query and reference (gallery)

images. These networks were first used in [40] for ReID that employ three feature extraction sub-networks for deep feature learning. Then, various deep learning architectures were proposed to learn discriminative feature embeddings. Most of these architectures [4, 5, 7, 8, 41, 47–49] employ end-to-end training, where both feature embedding and metric are learned as a joint optimization problem.

There are a number of metric learning losses that are widely used for optimizing deep ReID architectures. Contrastive loss is used in [4] to optimize a Siamese network that minimizes the distance between samples of the same class and forces a margin between samples of different classes. Triplet loss in ReID is first used in [47] that directly optimizes an embedding layer in Euclidean space which compares the relative distances of three training samples, namely an anchor image, a positive image sample from the same individual, and a negative sample from a different individual. In [7], original triplet loss is modified by adding an additional positive-pair constraint. A different version of triplet loss, named as quadruplet loss, is proposed in [5], which enlarges inter-class variations and reduces intra-class variation. Hermans et al. [41] extend the triplet loss by designing a simple semi-hard mining that selects the hardest positive and hardest negative of each anchor in a mini-batch. In [48], a soft hard-sample mining scheme is proposed by adaptively assigning weights to hard samples. On the other hand, magnet loss [49] is formulated as a negative log-likelihood ratio between the correct class and all other classes, but also forces a margin between samples of different classes. A thorough study of all the state-of-the-art metric learning losses for ReID suggests that triplet loss is the most widely used loss to optimize the deep ReID architecture. And among all the different versions of the triplet losses, the semi-hard mining-based triplet loss proposed by Herman et al. [41] is the simplest and efficient that does not require to change the backbone architecture to get the final feature embedding.

2.2 Multi-class classification loss

There has been an alternative trend that addresses the ReID problem as multi-class classification problems, where each ID is considered as a class. The objective of classification loss is to determine whether each input pair of images are the same or not, which makes full use of the ReID label with the predicted one from the classification networks. Some of the state-of-the-art classification-based ReID approaches [2, 38, 50] employ cross-entropy loss for image pairs in their network that takes pairwise images as inputs, and output the verification probability. Some other state-of-the-art ReID approaches [52, 53] used margin-based loss to keep the largest possible separation between the positive and negative pairs.

Recently, many works focus to learn local part-based feature which adopts the simple classification loss-based network performed on multiple local parts of a single image. Most of these approaches take into account the local features either from the human body part or by diving the global features to obtain discriminative part-based feature representations. State-of-the-art approaches [44–46] rely on diving the human body parts based on either external pose estimation or external semantic segmentation that leverage the semantic partitions for deeply learned part-based features. However, they highly depend on the efficiency of the external pose estimation or semantic segmentation techniques. In addition to that, they are suffering misalignment issues. Thus, to address these issues, state-of-the-art approaches [43] take into account global features and then divide into

parts or stripes. The advantage of using global features for local features representations is two folds: (i) does not suffer from misalignment caused by inaccurate bounding box detection, human pose changes, and various human spatial distributions and (ii) different channels of the global feature have different recognition patterns which increase the discriminative ability of the extracted feature by paying weighted attention to different parts of the human body. Thus, we are focusing on state-of-the-art methods that concentrate more on partitioning global features to form a local part-based feature representations.

2.3 Multiple losses

Recently, there have been efforts [54–57] to adopt multi-loss training strategies. More specifically, a combination of cross-entropy and triplet losses has proven to be effective to optimize ReID networks. The aim of this combination is to increase the discriminant power of the feature embedding by optimizing different objective functions. In [54], an omni-scale feature learning scheme is designed to capture the salient features at different scales that suggest optimizing the ReID network with the combination of cross-entropy and triplet losses for better performance. To achieve a similar objective of having multi-scale feature learning, Niki et al. [58] proposed a pyramid-inspired deep ReID architecture where multi-loss functions combined with curriculum learning strategy to optimize the network. In [59], an attention-driven Siamese learning architecture is designed to integrate attention and attention consistency by jointly optimizing the cross-entropy loss, the identification attention loss, and the Siamese attention loss. Chen et al. [60] use a reinforcement learning technique to quantify the attention quality and provide a powerful supervisory signal to guide the learning process. Following the same trend, they use the combination of cross-entropy and triplet losses to optimize their proposed architecture.

All of these ReID approaches above focus on improving the recognition accuracy without addressing the scalability issues to reduce computational costs. A few ReID approaches [61–66] seek to address the issues of computational complexity. Of these ReID approaches [61, 62], few rely on distillation-based approaches where knowledge is distilled from a deeper CNN (teacher model) to a lighter CNN (student model). Other approaches [63, 65, 66] rely on hashing to learn binary representation instead of real-value features for faster computation. In contrast to these approaches, our proposed ReID approach relies on the pruning techniques that compresses the deep CNN with a marginal reduction of recognition accuracy. Additionally, we propose different design scenarios or pipelines for leveraging a pruning method during the deployment of a CNN for a target ReID application domain.

3 Techniques for pruning CNNs

The objective of pruning is to remove unnecessary parameters from a neural network, while trying to maintain a comparable accuracy. Currently, pruning techniques operate on two different levels. First, techniques for weight-level pruning focus on pruning individual weights of a network. In contrast, techniques for channel-level pruning focus on removing all the parameters of the output and input channels of convolution layers. While weight pruning techniques can achieve high compression rate and good acceleration, its performance depends on a good sparse convolution algorithm which is unavailable and does not perform well on all platforms. In this paper, we focus on channel pruning techniques which do not rely on other algorithms and have been extensively studied in literature. This

section presents a survey of channel pruning techniques and summary of experimental results reported in the literature.

3.1 Channel pruning taxonomy

Table 2 presents the main properties of different pruning techniques according to strategy used to reduce channels. In order to facilitate the analysis of different pruning methods, we also categorize techniques according to the type of pruning criterion. In this table, “prune in one step” refers to techniques that prune the network one time and then fine-tune the network [33, 34, 67]. “Prune iteratively” is a type of pruning that is done iteratively on a trained model that alternates between pruning and fine-tuning [32]. Pruning by regularization is usually done by adding a regularization term to the original loss function in order to leave the pruning process for the optimization [69, 70]. Pruning by minimizing the reconstruction error is a family of algorithms that tries to minimize the difference of outputs between the pruned and the original model. “Progressive pruning,” while very similar to iterative pruning, differs in that it can start directly from a model that was not trained and progressively prune it during training.

One key challenge of pruning neural networks is selecting the pruning criteria. It should allow to discern the parameters that contribute to accuracy and the ones that do not. Another challenge is finding an optimal pruning compression. This compression ratio is essential to find a compromise between the reduction of complexity for the model and the loss of accuracy. Finally, one challenge is the retraining and pruning schedule of the model. Pruning can be performed in one iteration but the damage caused to the network may be considerable. On the counterpart, we could prune and retrain iteratively to reduce this damage at each iteration, but this will take longer to apply. The retraining of the pruned network may also cause the model to overfit or get caught in local minimums.

3.2 Description of methods

This subsection presents different pruning algorithms for each pruning family in the taxonomy. To ease our notation, we refer to a convolution tensor as \mathbf{W} with $\mathbf{W} \in$

Table 2 Main properties of different channel pruning techniques

Strategy	Methods	Criteria
Prune in one step	L1 [33]	Weights: $S_j = \sum w_k $
	Redundant channels [67]	Weights: $SIM_C(\mathbf{W}_i, \mathbf{W}_j) = \frac{\mathbf{w}_i \cdot \mathbf{w}_j}{\ \mathbf{w}_i\ \cdot \ \mathbf{w}_j\ }$
	Entropy [34]	Feature maps: $E_j = -\sum_{a=1}^m (p_a \log(p_a))$
Prune iteratively	Taylor [32]	Feature maps: $ \Delta C(\mathbf{H}_{ij}) = \left \frac{\delta C}{\delta \mathbf{H}_{ij}} \mathbf{H}_{ij} \right $
	FPGM [68]	Weights: $\mathbf{W}_{ij}^* \in \underset{\mathbf{w} \in \mathbb{R}^{n_m \times k \times k}}{\operatorname{argmin}} \sum_{j' \in [1, n_{out}]} \ x - \mathbf{w}_{ij'}\ _2$
Prune iteratively with regularization	Play and Prune [69]	Weights: $S_j = \sum w_k $
	Auto-Balance [70]	Weights: $S_j = \sum w_k $
Prune iteratively, min reconstruction error	ThiNet	Feature maps: $\mathbf{H}_{i+1,j} = \sum_{j=1}^C \sum_{k=1}^K \sum_{k=1}^K \mathbf{W}_{ij,k,k} * \mathbf{H}_{ij}$
	Channel pruning [35]	Feature maps: $\arg \min_{\beta, \mathbf{W}} \frac{1}{2N} \left\ \mathbf{H}_{i+1,j} - \sum_{j=1}^n \beta_{ij} \mathbf{H}_{ij} \mathbf{W}_{ij} \right\ _F^2 + \lambda \ \beta\ _1$
Prune progressively	PSFP [36]	Weights: $S_j = \sum w_k _2$

$\mathbb{R}^{n_{out} \times n_{in} \times k \times k}$, n_{in} the number of input channels, n_{out} the number of output channels, and k the kernel size. An output channel tensor i is then defined as \mathbf{W}_i , and an individual weight is defined as \mathbf{w} . For feature map, \mathbf{H} represents an output of a convolution layer and \mathbf{H}_i then represents the output channel of a feature map. For ease of notation, we do not mention the layer index unless necessary; therefore, \mathbf{W} or \mathbf{H} can be any convolution layer or feature map at any index.

3.2.1 Criteria based on weights

The L1 [33] pruning algorithm is a layer-by-layer method which means it will prune the network one layer at a time. This algorithm’s pruning criteria are simple and could be implemented using Algorithm 1. The retraining could be done in two different ways:

Algorithm 1 L1 Pruning

- 1: **Input:** training data: \mathbf{X}
 - 2: **Input:** pruning rate: P_i
 - 3: **Input:** the model with parameters $\mathbf{M} = \{\mathbf{M}^{(i)}, 0 \leq i \leq L\}$
 - 4: Initialize the model parameter \mathbf{M}
 - 5: **for** each convolution layer **do**
 - 6: Calculate l_1 -norm for each channel
 - 7: Select the N lowest l_1 -norm depending on the pruning rate
 - 8: Remove the N selected channels
 - 9: **end for**
 - 10: Re-Train the pruned network
 - 11: **Output:** Compact model with parameters M'
-

- 1 Prune once over multiple layers and retrain (more adapted for resilient layers)
- 2 Prune channels one by one and retrain each time (more adapted for layers that are less resilient)

For this algorithm, we chose to mix the two retraining methods to come up with pruning N channels before retraining.

The second weight method that will be presented is the redundant channel pruning [67]. This method’s idea is to pruned channels that are similar to the ones that are kept. To do so, the authors proposed to regroup each channel of a layer in n_f clusters depending on a similarity score being higher than a preassigned threshold τ . To determine the similarity between these channels, the authors proposed to use the cosine similarity between the weights of the channels.

$$\overline{SIM_C}(C_a, C_b) = \frac{\sum_{\mathbf{w}_i \in C_a, \mathbf{w}_j \in C_b} SIM_C(\mathbf{W}_i, \mathbf{W}_j)}{|C_a| \times |C_b|} > \tau \tag{1}$$

$a, b = 1, \dots, n_{out}; a \neq b; i = 1, \dots, |C_a|; j = 1, \dots, |C_b|; i \neq j$

With the calculation of SIM_C of two output channel given below:

$$SIM_C(\mathbf{W}_i, \mathbf{W}_j) = \frac{\mathbf{W}_i \cdot \mathbf{W}_j}{\|\mathbf{W}_i\| \cdot \|\mathbf{W}_j\|} \tag{2}$$

Equation 2 gives us the ability to determine the similarity between two channels by calculating the cosine of the angle between two vectors of dimension n . The pruning of one specific layer could be done in 2 steps:

- 1 Group the channels in the same cluster if $\cos(\theta)$ from Eq. 1 is above the threshold τ
- 2 Randomly sample one channel in each cluster and pruned the remaining ones of each cluster.

The threshold τ acts as the compression ratio in this pruning algorithm where a low threshold means a high compression rate and vice versa (see Algorithm 2).

Algorithm 2 SIM_C Pruning

- 1: **Input:** training data: \mathbf{X}
 - 2: **Input:** pruning threshold τ
 - 3: **Input:** the model with parameters $\mathbf{M} = \{\mathbf{M}^{(i)}, 0 \leq i \leq L\}$
 - 4: Initialize the model parameter \mathbf{M}
 - 5: **for** each convolution layer **do**
 - 6: Calculate the similarity score for each pair of channel
 - 7: Separate the channels into two clusters based on threshold τ
 - 8: Select the N lowest l_1 -norm depending on the pruning rate
 - 9: Randomly sample one channel in each cluster and pruned the remaining ones of each cluster.
 - 10: **end for**
 - 11: Re-Train the pruned network
 - 12: **Output:** Compact model with parameters M'
-

The third weight-based method that will be presented is the Auto-Balanced pruning [70] that uses the same pruning criteria as the L1 algorithm which is a L1 norm of weight kernels to determine the ranking of the channels. But this method adds a regularization term during the training to transfer the representational capacity of the channels we want to prune to the remaining ones. In order to calculate this transfer of representational capacity, the authors proposed to separate the channels in two subsets at the beginning of each pruning iteration. In order to assign the channels to their subset, the authors used the L1 norm of the weights of the channels. The vec function is used to flatten the weight matrix into a vector and $\mathbf{M}_{i,j}$ the metric measuring the importance. Here, we use the notation of $\mathbf{W}_{i,j}$ with i representing the layer index and j the output channel index.

$$\mathbf{M}_{i,j} = \|\mathit{vec}(\mathbf{W}_{i,j})\|_1 \tag{3}$$

Once the L1 score has been calculated for each channel, they are then assigned to one of the subsets depending on the threshold θ which is fixed depending on the desired number of remaining channels per layer.

$$\mathbf{M}_{i,j} > \theta_i \forall \mathbf{W}_{i,j} \in \mathbf{P}_i \tag{4}$$

$$\mathbf{M}_{i,j} < \theta_i \forall \mathbf{W}_{i,j} \in \mathbf{R}_i \tag{5}$$

The channels in subset \mathbf{R} (remaining) and subset \mathbf{P} (to pruned) are then adjusted with an L2 regularization term. The following equations are used to calculate this L2 adjustment factor:

$$\lambda_{i,j} = \begin{cases} 1 + \log \frac{\theta_i}{\mathbf{M}_{i,j} + \varepsilon} & \text{if } \mathbf{W}_{i,j} \in \mathbf{P}_i \\ -1 - \log \frac{\mathbf{M}_{i,j}}{\theta_i + \varepsilon} & \text{if } \mathbf{W}_{i,j} \in \mathbf{R}_i \end{cases} \quad (6)$$

$$S(\mathbf{P}_i) = \sum_{\mathbf{W}_{i,j} \in \mathbf{P}_i} \lambda_{i,j} \|\text{vec}(\mathbf{W}_{i,j})\|_2^2, \quad S(\mathbf{R}_i) = \sum_{\mathbf{W}_{i,j} \in \mathbf{R}_i} \lambda_{i,j} \|\text{vec}(\mathbf{W}_{i,j})\|_2^2 \quad (7)$$

$$S(\mathbf{P}) = \sum_{i=1}^n S(\mathbf{P}_i), \quad S(\mathbf{R}) = \sum_{i=1}^n S(\mathbf{R}_i) \quad (8)$$

The cost function for training is changed with Eq. 9 where L_0 represents the original cost function.

$$L = L_0 + \alpha S(\mathbf{P}) + \tau S(\mathbf{R}) \quad (9)$$

$$\tau = -\alpha \frac{S(\mathbf{P})}{S(\mathbf{R})} \quad (10)$$

This enables the model to penalize the weak channels and stimulate the strong ones. This method adds two hyper-parameters in the training which are α and r . α is the regularization factor and the vector r is the target of remaining channels in each layer (see Algorithm 3).

Algorithm 3 Auto-Balance

- 1: **Input:** training data: \mathbf{X}
 - 2: **Input:** pruning threshold τ, α, r
 - 3: **Input:** the model with parameters $\mathbf{M} = \{\mathbf{M}^{(i)}, 0 \leq i \leq L\}$
 - 4: Initialize the model parameter \mathbf{M}
 - 5: **for** each convolution layer **do**
 - 6: Divide the channels into R (remain) subset and P (Prune) subset using l_1 -norm
 - 7: Optimize the Equation 10
 - 8: **end for**
 - 9: **Output:** Compact model with parameters M'
-

The last weight-based method is the progressive soft pruning [36] where their pruning criterion is the same as the L1 method (L2 norm of the weights). The main difference with this method is they proposed an interesting pruning scheme that allows pruning during the fine-tuning step. The authors proposed to use soft pruning which means instead of removing the channels during the pruning, they set the weights to 0 and allow these channels to be updated during the retraining phase. This pruning scheme is very interesting since the model keeps its original dimension during the retraining phase. The authors also proposed to add a progressive pruning scheme where at each pruning iteration, the compression ratio is increased in order to get a shallower network. Once these iterations of pruning and retraining are completed, they do a last channel ranking using a pruning criteria and they discard the lowest channels depending on the compression ratio. Their pseudo-code for the progressive soft pruning scheme can be viewed in Algorithm 4. In

Algorithm 4 Algorithm Description of PSFP

- 1: **Input:** training data: X
 - 2: **Input:** pruning rate: P_i , pruning rate decay D
 - 3: **Input:** the model with parameters $M = \{M^{(i)}, 0 \leq i \leq L\}$
 - 4: Initialize the model parameter M
 - 5: **for** $epoch = 1; epoch \leq N_{epoch}; epoch ++$ **do**
 - 6: Update the model parameters M based on X
 - 7: **for** $i = 1; i \leq L; i ++$ **do**
 - 8: Calculate the l_2 -norm for each channel
 - 9: Calculate the pruning rate P' at this epoch using P_i and D
 - 10: Select the N lowest l_2 -norm depending on the pruning rate
 - 11: Zeroize the weights W of the selected channels
 - 12: **end for**
 - 13: **end for**
 - 14: Obtain the compact model with parameters M' from M
 - 15: **Output:** Compact model with parameters M'
-

the article, they used the L1 or L2 norm of the weights as pruning criteria which means this method could be categorized as a weight-based method.

The L represents the number of layers in the model, i represents the layer number, W represents the weights of a channel, and N is the number of channels to prune. The pruning rate P' is calculated at each epoch using the pruning rate goal P_i for the corresponding layer i and the pruning rate decay D . To calculate the pruning rate, we can use Eq. 11

$$P'_i = a \cdot e^{-k \cdot epoch} + b \tag{11}$$

The a, b , and k values can be calculated by solving Eq. 12

$$\begin{cases} 0 = a + b \\ \frac{P_i}{4} = e^{-k \cdot N_{epoch} \cdot D} + b \end{cases} \tag{12}$$

FPGM [68] is a new technique that focuses on using a geometric median to prune away output channels. A geometric median is defined as follows: given a set of n points $A = [a_1, a_2, \dots, a_n]$ with $a_i \in R^d$, find a point $x^* \in R^d$ that minimizes the sum of the Euclidean distances to them:

$$x^* \in \underset{x \in R^d}{\operatorname{argmin}} f(x) \text{ where } f(x) \triangleq \sum_{i \in [1, n]} \|x - a_i\|_2 \tag{13}$$

Using Eq. 13, a geometric median F_i^{GM} for all the filters of a layer i can be found:

$$\begin{aligned} \mathbf{W}_i^{GM} &\in \underset{x \in R^{n_{out} \times k \times k}}{\operatorname{argmin}} g(x) \\ \text{where } g(x) &\triangleq \sum_{j \in [1, n_{in}]} \|x - \mathbf{W}_{i,j}\|_2 \end{aligned} \tag{14}$$

In order to select, non-important output channels, the author proposed to find the channels that have the same or similar value of \mathbf{W}_i^{GM} which translates to:

$$\mathbf{W}_{i,j^*} \in \underset{x \in R^{n_{out} \times k \times k}}{\operatorname{argmin}} \|\mathbf{W}_{i,j^*} - \mathbf{W}_i^{GM}\|_2 \tag{15}$$

Since geometric median is a non-trivial problem, it is quite computationally intensive; therefore, the authors propose to relax the problem by assuming that:

$$\|\mathbf{W}_{i,j^*} - \mathbf{W}_i^{GM}\|_2 = 0 \tag{16}$$

This transforms Eq. 14 to:

$$\begin{aligned} \mathbf{W}_{i,j^*} \in \operatorname{argmin}_{j^* \in \mathcal{R}^{n_{in} \times k \times k}} \sum_{j' \in [1, n_{out}]} \|\mathbf{x} - \mathbf{W}_{i,j'}\|_2 \\ = \operatorname{argmin}_{j^* \in \mathcal{R}^{n_{in} \times k \times k}} g(\mathbf{x}) \end{aligned} \tag{17}$$

The algorithm of FPGM is summarized in Algorithm 5.

Algorithm 5 Algorithm Description of FPGM

- 1: **Input:** training data: \mathbf{X}
 - 2: **Input:** pruning rate: P
 - 3: **Input:** the model with parameters $\mathbf{M} = \{\mathbf{M}^{(i)}, 0 \leq i \leq L\}$
 - 4: Initialize the model parameter \mathbf{M}
 - 5: **for** $epoch = 1; epoch \leq N_{epoch}; epoch++$ **do**
 - 6: Update the model parameters \mathbf{M} based on \mathbf{X}
 - 7: **for** $i = 1; i \leq L; i++$ **do**
 - 8: Select the $n_{out} \times P$ of W_i channels that satisfy Equation 17
 - 9: Set the selected channels to zero
 - 10: **end for**
 - 11: **end for**
 - 12: Obtain the compact model with parameters \mathbf{M}' from \mathbf{M}
 - 13: **Output:** Compact model with parameters M'
-

Play and Prune [69] is an adaptive output channel pruning technique that, instead of focusing on a criterion, tries to find an optimal number of output channels that can be pruned away given an error tolerance rate. This technique is a min-max game of two modules, The Adaptive Filter Pruning (AFP) module and the Pruning Rate Controller (PRC). The goal of the AFP is to minimize the number of output channels in the model while the PRC tries to maximize the accuracy of the remaining set of output channels. This technique considers a model M can be partitioned into two sets of important channels \mathbf{I} and unimportant channels \mathbf{U} .

$$U_i = \sigma_{\text{top } \alpha\%}(\text{sort}(\{|\mathbf{W}_1|, |\mathbf{W}_2|, \dots, |\mathbf{W}_{n_{out}}|\})) \tag{18}$$

U_i represents all the unimportant channels of a layer i . It is selected by selecting $\alpha\%$ channels of the result of the sort operation on the L1 norm of each output channel. Once an U_i is selected, the authors propose to add an additional penalty to the original loss function in order to prune without loss of accuracy while helping the pruning process. The original loss function would then become:

$$\Theta = \operatorname{argmin}_{\Theta} (C(\Theta) + \lambda_A \|\mathbf{U}\|_1) \tag{19}$$

where $C(\Theta)$ is the original cost function to optimize the original model parameters, and λ_A is the L_1 regularization term. While this optimization helps pushing the channels to

have zero sum of absolute weights, it can take some epochs; therefore, the authors propose an adaptive weight threshold (\mathbf{W}_i) for each layer i . Any channels with L1 norm below this threshold will be removed. While this value is given by the PRC, for the first epoch, it is found by using a binary search on the histogram of sum of absolute weights. The AFP minimizes the number of output channels in the model using Eq. 19. The AFP can be summarized in Algorithm 6, and the loss function of AFP can be written as:

$$\Theta' = \sigma_{\#w \in \Theta'} \left[P \left(\underset{\Theta'}{\operatorname{argmin}} (C(\Theta) + \lambda_A \|\mathbf{U}\|_1) \right) \right] \tag{20}$$

For the PRC, the adaptive threshold W_A is updated as follows:

$$\mathbf{W}_A = \delta_w \times T_r \times \mathbf{W}'_A \tag{21}$$

with δ_w the constant used to increase or decrease the pruning rate. T_r is calculated as follows:

$$T_r = \begin{cases} C(\#w) - (\xi - \varepsilon) : C(\#w) - (\xi - \varepsilon) > 0 \\ 0 : \text{Otherwise} \end{cases} \tag{22}$$

where ξ is the accuracy of the unpruned network, ε is the tolerance error, and $C(\#w)$ is the accuracy of the model with the remaining filter $\#w$. The regularization constant is also computed as follows:

$$\lambda_A = \begin{cases} C(\#w) - (\xi - \varepsilon) \times \lambda : C(\#w) - (\xi - \varepsilon) > 0 \\ 0 : \text{Otherwise} \end{cases} \tag{23}$$

with λ the initial regularization constant. By alternating between the AFP and the PRC, the authors propose a system that prunes at each epoch in an adaptive and iterative way (see Algorithm 6).

Algorithm 6 AFP

- 1: **Input:** training data: \mathbf{X}
 - 2: **Input:** the model with parameters $\mathbf{M} = \{\mathbf{M}^{(i)}, 0 \leq i \leq L\}$
 - 3: Initialize the model parameter \mathbf{M}
 - 4: **for** each convolution layer **do**
 - 5: Select an $\alpha\%$ of output channels with the lowest l_1 -norm
 - 6: Separate the output channels into U (Unimportant) and I (Important) subsets
 - 7: Perform Equation 19 with λ_A Agiven by PRC
 - 8: Remove unimportant channels using the threshold W_A given by PRC
 - 9: **end for**
 - 10: **Output:** Compact model with parameters M'
-

3.2.2 Criteria based on feature maps

In the channel-based approach, we are going to present 3 algorithms which are ThiNet [71], Channel Pruning [35], and Entropy Pruning [34]. The first two algorithms have the same idea behind their pruning algorithm which is minimizing the difference in the activation maps but they diverge with their minimization technique. ThiNet’s goal is to find a subset of channels that minimize the difference in the output at layer $i+1$ (feature map).

ThiNet uses greedy algorithm to find which subset of channels to eliminate and keep the input at layer $i+2$ almost intact. To find the subset of channels to prune, the authors

proposed to use a greedy algorithm where they compute the value for each channel in a layer and assign the lowest value to the subset. They repeat this method until our pruning subset respects the defined compression ratio. To calculate the input of the feature map in layer $i+2$, we can use Eq. 24.

$$\mathbf{H}_{i+1,j} = \sum_{j=1}^C \sum_{k=1}^K \sum_{k=1}^K \mathbf{W}_{i,j,k,k} * \mathbf{H}_{i,j} \tag{24}$$

where i represents the layer, j the channel index, and k the kernel size of the channel. To compute the value of a channel, the authors proposed to used Eq. 25

$$\sum_{i=1}^m \hat{x}_{i,j}^2 \tag{25}$$

where \hat{x} is equal to $\mathbf{W}_{i+1,j}$ in Eq. 24. This greedy method is repeated for each layer needed to be pruned in the model.

The Channel Pruning method also has the goal to minimize the difference in the output (feature map) but their method is to find a subset of channels with a LASSO regression.

$$\underset{\beta, \mathbf{W}}{\operatorname{argmin}} \frac{1}{2N} \left\| \mathbf{H}_{i+1,j} - \sum_{j=1}^n \beta_{i,j} \mathbf{H}_{i,j} \mathbf{W}_{i,j} \right\|_F^2 + \lambda \|\beta\|_1 \tag{26}$$

$\|\beta\|_0 \leq n'$
 $0 \leq n' \leq n$

β represents channel mask that decides whether the channel is pruned or not. If β is zero, then the channel is no longer useful. The compression ratio is defined with λ . The n represents the number of channels and n' represents the number of remaining channels. During the pruning iterations, the W in Eq. 26 is fixed which leaves us with only one variable to minimize which is β . The LASSO regression is used to find this β mask that minimizes the difference in the output. As in the ThiNet method, this method also requires to redo these steps for every layer needed to be pruned.

The entropy pruning [34] method is also a layer-by-layer algorithm but instead of trying to minimize the difference in the output like the two channel methods above, they used a different criteria based on the entropy of the feature maps produced by the channels. The idea behind their criteria is that a low entropy in the feature maps of a channel will most likely be less important in the decision of the network. With the entropy criterion defined as:

$$E_j = - \sum_{a=1}^m (p_a \log(p_a)) \tag{27}$$

Pruning for layer i is done according to Algorithm 7.

Taylor's [32] pruning algorithm seeks to minimize the cost function. It approximates the change in this function if the channel is pruned according to:

$$|\Delta C(\mathbf{H}_{i,j})| = |C(D, \mathbf{H}_{i,j} = 0) - C(D, \mathbf{H}_{i,j})| \tag{28}$$

where C represents the cost function and D the dataset. $C(D, \mathbf{H}_{i,j} = 0)$ is the cost value if channel $\mathbf{H}_{i,j}$ is pruned. The idea is to find a subset of channels $H_{i,j}$ to prune while minimizing the difference with the original cost function were these channels were used. This is represented in the equation by calculating the difference between cost function with the channels excluded and the cost function with the channels included. Using a Taylor

Algorithm 7 Entropy Pruning

- 1: **Input:** training data: \mathbf{X}
 - 2: **Input:** Pruning rate: \mathbf{P}
 - 3: **Input:** the model with parameters $\mathbf{M} = \{\mathbf{M}^{(i)}, 0 \leq i \leq L\}$
 - 4: Initialize the model parameter \mathbf{M}
 - 5: Run all the data through the network and collect features for each convolution layer
 - 6: **for** each convolution layer **do**
 - 7: Convert the activation maps into a vector of dimension n_{out} (number of channels) using a global average pooling.
 - 8: For each channel j divide the distribution into m bins and calculate the entropy using 27
 - 9: Remove channels with the lowest entropy value according to the pruning rate
 - 10: **end for**
 - 11: **Output:** Compact model with parameters M'
-

expansion to solve this minimization, the authors found that the difference in the cost function with the channels pruned could be approximated with the activation (feature map) and the gradient of the channel which can be calculated during back-propagation.

$$|\Delta C(\mathbf{H}_{i,j})| = \left| \frac{\delta C}{\delta \mathbf{H}_{i,j}} \mathbf{H}_{i,j} \right| \quad (29)$$

Each channel ranking value is normalized using a l2-normalization. This normalization is done on each layer individually in order to facilitate the comparison between layers since this method ranks channels across all layers (see Algorithm 8):

Algorithm 8 Taylor Pruning

- 1: **Input:** training data: \mathbf{X}
 - 2: **Input:** Stopping condition
 - 3: **Input:** the model with parameters $\mathbf{M} = \{\mathbf{M}^{(i)}, 0 \leq i \leq L\}$
 - 4: Initialize the model parameter \mathbf{M}
 - 5: **for** $epoch = 1; epoch \leq N_{epoch}; epoch++$ **do**
 - 6: Evaluate the importance of output channels using X and Equation 29
 - 7: **for** each convolution layer **do**
 - 8: Remove channel with the least importance
 - 9: **end for**
 - 10: Fine-tune the pruned network
 - 11: **if** Stopping condition is **True**
 - 12: **Break**
 - 13: **end for**
 - 14: Obtain the compact model with parameters M' from \mathbf{M}
 - 15: **Output:** Compact model with parameters M'
-

3.2.3 Output-based

The second output-based method is the Neuron Importance Score Propagation (NISIP) [72] where the pruning is done by back-propagating channel scores across the model to

determine which ones to prune. The intuition behind their idea is to use a feature ranking method on the last layer before the classification since this layer is the one to play a more significant role in our application. Once every feature has an associated score, the authors propose to back-propagate that score into the network to have an importance score for each channel in the network. The importance score is then used to determine which channels we pruned and which one we retain by using a predefined compression ratio for each layer in the model.

The score is back-propagated using Eq. 30.

$$S_{i,j} = \sum_i |W_{i,j}^{(k+1)}| S_{i+1,j} \quad (30)$$

where W is the weights, j is the neuron or channel, i is the layer, and k is the number of connections from that neuron to the next layer. This equation represents a weighted sum of the scores in the subsequent layers.

3.3 Critical analysis of pruning methods

The main difference between methods using a weight-based criteria versus methods based on feature map criteria is that they are not dependent on a dataset since weight statistics do not depend on output of a CNN. Methods based on feature maps need a dataset in order to compute the output of convolution layers or its gradients.

The chosen criteria usually depend on the desire to simplify the pruning steps at the expense of a lower accuracy—some more complex criteria that require more computations allow preserving a high level of accuracy. If training and pruning time is an issue, i.e., applications with design constraints, and that requires fast deployment, simple criteria like L1 and L2 norm are more suitable. However, if there are no complexity constraints, some more complex pruning criteria, like the minimization in the difference of activation or cost functions, can outperform the simpler criteria, at the expense of more computations and time.

Some of the techniques also differ in terms of how channels are pruned, some prune layer-by-layer [33, 36], while others prune across layers [32]. One of the differences between across layer and layer-by-layer pruning is the imbalance in terms of pruning. An across layer pruning does not prune each layer evenly, and the method can possibly prune lower level layers more than higher level layers, and vice versa. Depending on the CNN architecture and pruning algorithm, pruning across layers may not yield the desired reduction. Layer-by-layer pruning can guarantee that all the layers will be pruned and therefore undergo a more even reduction at each layer.

Recently, some techniques [36, 68] also adopt a new soft pruning approach. Soft pruning differs from hard pruning because it only resets pruned channels to zero instead of completely removing them. Therefore, soft pruned channels have a chance to recover. These techniques have been shown to have achieved state-of-the-art performance.

Table 3 summarizes a comparative experimental analysis of different pruning techniques. All the results reported in this table have been taken from the corresponding papers. Experimental performance indicates that VGG16 processing can accelerate by up to 2.5 times at the expense of increased error of 1% on the ImageNet dataset. Comparing L1 [33] versus Auto-Balanced [70] techniques, both based on a weight-based criteria, we observe that the Auto-Balanced techniques can obtain higher compression ratios because

Table 3 Performance of pruning techniques from the literature in terms of rank-1 accuracy and computational complexity (memory(M): number of parameters and time (T): time required for one forward pass). To ease comparison, we include the out results produced with ThiNet (channel pruning method)

Dataset		ResNet56 trained on CIFAR10				
Algorithm	Original			Pruned		
	rank-1	T	M	rank-1	T	M
L1 [33]	93.04	0.125	0.85	93.06	0.091	0.73
Auto-Balanced [70]	93.93	0.142	N/D	92.94	0.055	N/D
Redundant channel [67]	93.39	0.125	0.85	93.12	0.091	0.65
Play and Prune [69]	93.39	0.125	0.85	93.09	0.039	N/D
FPGM [68]	93.39	0.125	0.85	92.73	0.059	N/D

Dataset		VGG16 trained on ImageNet				
Algorithm	Original			Pruned		
	rank-1	T	M	rank-1	T	M
ThiNet [71]	90.01	30.94	138.34	89.41	9.58	131.44
Taylor [32]	89.30	30.96	N/D	87.06	11.5	N/D
HaoLi [33]	90.01	30.94	138.34	89.13	9.58	130.87
Channel Pruning [35]	90.01	30.94	138.34	88.10	7.03	131.44

Dataset		ResNet50 trained on ImageNet				
Algorithm	Original			Pruned		
	rank-1	T	M	rank-1	T	M
Entropy [34]	72.88	3.86	25.56	70.84	2.52	17.38
ThiNet [71]	75.30	7.72	25.56	72.03	3.41	138.00
FPGM [68]	75.30	7.72	25.56	74.83	3.58	N/D

of their regularization term. If we compare weight-based and channel-based approaches, we observe that using channel-based provides a higher compression while maintaining similar accuracy. For the comparison between output and channel-based approaches, ThiNet [71] outperforms Taylor [32] in accuracy and complexity.

3.4 Design scenarios with pruning

Most DL models for person ReID use pretraining, and then fine-tune the model to the task or target application domain. Pretraining is typically performed using a large-scale dataset in order to prime CNN parameters towards relevant optimization solutions. In many cases, CNNs are pretrained on ImageNet since this public dataset has a large amount of diverse training samples from different classes which improves the CNN capacity to generalize. In person ReID, pretrained models have proven to be more successful than models that were trained from scratch directly on the task dataset.

Once the model has been pretrained, the next step is fine-tuning to map the model’s parameters from our pretraining source domain to our target application domain. It is crucial that the task dataset be similar to pretraining data. As described in [73], the best fine-tuning practices depend on the size of the task training dataset, and the difference in data distribution between pretraining and task domain data. The authors propose to compute a similarity score between the pretraining and task datasets in order to guide the fine-tuning from one target domain to another. They proposed measuring their similarity with the cosine distance and the maximum mean discrepancy (MMD). In particular, they proposed to average the feature embedding of each dataset and calculate the metrics between the two vectors. Given these metrics and the number of samples per class in the

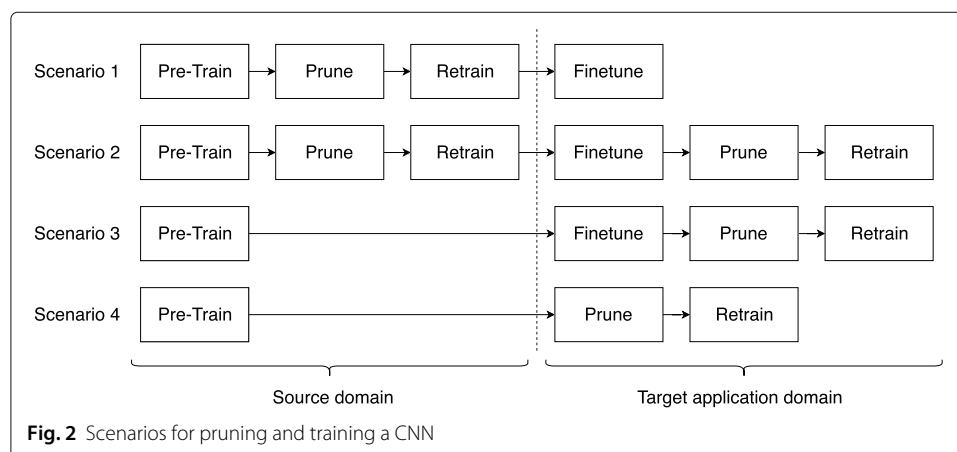
target domain dataset, authors proposed to either train the whole network or freeze the feature extractor and fine-tune the classifier. We follow their guidelines to determine the layers to freeze and to fine-tune.

Pruning neural networks can be done in both main training phases—pretraining and fine-tuning. We concluded that there are four possible scenarios for pruning (as shown in Fig. 2). The first scenario consists in pruning a CNN on the source pretraining dataset. The idea behind this scenario is to leverage a large-scale dataset to guide our selection of the more relevant and discriminant source domain channels. The second scenario consists in pruning on the source pretraining dataset, and then fine-tuning until our model provides a suitable performance, and then prune again on the target application dataset. This strategy allows removing additional channels that are not contributing to our task. The third scenario consists in only pruned on our task dataset after the fine-tuning on the target application dataset. The objective of this scenario is to accelerate the training time since pruning and retraining on a large-scale source domain dataset can be time consuming. Finally, the last scenario consists in pruning on the task dataset before doing the fine-tuning. This scenario goal also reduces design time of the model. In Section 5, we seek to determine the best scenario to reduce the computational complexity of CNNs, while maintaining a comparable level of accuracy on our task.

The progressive soft pruning method is an interesting alternative since the model is pruned during fine-tuning steps. This pruning scheme can reduce training effort since it combines the pruning, retraining, and fine-tuning into a single step. In Fig. 2, progressive soft pruning would be represented by combining the prune and retrain process in one box for Scenario 1. For Scenarios 2 and 3, the fine-tuning, pruning, and retraining would be combined into one box. As for Scenario 4, PSFP is not applicable since the pruning, retrain, and fine-tuning is one step, making it impossible to prune the network by ranking the channels with the target data and then fine-tuning the network.

4 Experimental methodology

In this section, we present the experimental methodology used to validate the pruning model. Our experiment is divided into two main parts. First, we experiment on a large-scale dataset, i.e., ImageNet, in order to find the best pruning methods using the same experimental protocol. The second part of these experiments will be to test the



pruning algorithms on a person ReID problem to find the advantage of using a pruned model compared to a smaller model. The following section will present the experimental methodology such as the datasets, the evaluation metrics, and the experiment algorithm. The results for the pruning on the ImageNet dataset and ReID datasets will also be presented.

4.1 Datasets

Four publicly available datasets are considered for the experiments, namely Imagenet [74], Market1501 [37], DukeMTMC-reID [39], and CUHK03-NP [38]. Imagenet, a large-scale dataset, is used as a pretrained dataset and the rest of the other datasets (small-scale) are used for the experiments of person ReIDs.

- **ImageNet** (ILSVRC2012) [74] is composed of two parts. The first part is used for training the model and the second part is used for validation/testing. There is 1.2M images for training and 50k for validation. The ILSVRC2012 dataset contains 1000 classes of natural images.
- **Market-1501** [37] is one of the largest public benchmark datasets for person ReID. It contains 1501 identities which are captured by six different cameras, and 32,668 pedestrian image bounding boxes obtained using the Deformable Part Models (DPM) pedestrian detector. Each person has 3.6 images on average at each viewpoint. The dataset is split into two parts: 750 identities are utilized for training and the remaining 751 identities are used for testing. We follow the official testing protocol where 3368 query images are selected as a probe set to find the correct match across 19,732 reference gallery images.
- **CUHK03-NP** [38] consists of 14,096 images of 1467 identities. Each person is captured using two cameras on the CUHK campus and has an average of 4.8 images in each camera. The dataset provides both manually labeled bounding boxes and DPM-detected bounding boxes. In this paper, both experimental results on “labeled” and “detected” data are presented. We follow the new training protocol proposed in [75], similar to partitions of the Market1501 dataset. The new protocol splits the dataset into training and testing sets, which consist of 767 and 700 identities, respectively.
- **DukeMTMC-reID** [39] is constructed from the multi-camera tracking dataset—DukeMTMC. It contains 1812 identities. We follow the standard splitting protocol proposed in [76] where 702 identities are used as the training set and the remaining 1110 identities as the testing set.

4.2 Pruning methods

For our experiments, we compare five pruning methods in order to determine which technique gives the best compression ratio while maintaining a good performance on person ReID task. Our choice was based on the following criteria: article results, most of the families of the taxonomy are represented and the complexity for the ranking and the implementation. We selected `L1` [33] and `Entropy` [34] as they rely on the techniques that prune the network only one time and then fine-tune the network. Although `Taylor` [32] uses iterative pruning techniques, we chose this method for our experiments because of its theoretical explanation and requires a single compression ratio. We choose to experiment with `Auto-Balanced` algorithm [70] because pruning is done

by adding regularization terms to the original loss function in order to leave the pruning process for the optimization. We have also decided to try the `Progressive Soft Pruning` [36] method since it directly prunes from scratch and progressively prunes during training which is a suitable test on our target operational domain.

4.2.1 Implementation details

For the Triplet-based ReID method, images are resized to 256×128 for all the datasets. For PCB [43] architectures, images are resized to 384×128 . Like many state-of-art ReID approaches [3, 5–8, 43], we use ResNet50 [11] as the backbone architecture, where the final layer is removed to get a 2048 feature representation. We apply all the pruning methods on the ResNet50 architecture. In order to be able to compare the four algorithms more easily, we decided to come up with a pruning schedule that would be similar for all the methods. First of all, we decided to prune around 5% of the total number of channels at each iteration.

For the layer-by-layer methods, we chose to use a single compression rate for every layer in order to simplify our experiments and our comparison between the methods. For each pruning iteration, we decided to use 1 epoch for the ranking of the channels and 4 epochs for retraining before moving to the next iteration.

This pruning schedule was used for every method on ImageNet in order to produce our pruned models that would be used in the person ReID experiments. We have discarded the pruning iterations where the accuracy was too low since there was no advantage of using these networks for our task. Once our pruning was done for every method, we retrained every model on ImageNet to regain the loss of accuracy caused by the pruning. Each of our pruned models was then fine-tuned on the ReID datasets. We also fine-tuned pretrained ResNet18 and ResNet34 on these ReID datasets in order to compare the advantages of using pruned models compared to shallower networks.

4.3 Performance metrics

Following the common trend of evaluation [3, 5–8], we use the rank-01 accuracy of the cumulative matching characteristics (CMC) and the mean average precision (mAP) to evaluate the ReID accuracy. The CMC represents the expectation of finding a correct match in the top n ranks. When multiple ground truth matches are available, then CMC cannot measure how well the gallery images are ranked. Thus, we also report the mAP scores.

As the state-of-the-art pruning methods [33–36], the FLOPS's metric is used to calculate the model's complexity in terms of computational operations. To compare the different models during our experiments, we decided to calculate the number of FLOPS necessary to process one image through the model. We chose to compare the number of FLOPS since the processing time depends on the material used. The FLOPS is also a better metric than the number of pruned channels since a pruned channel at the beginning of the network will be reduced considerably more the total number of FLOPS than a later layer channel since the image dimension is reduced throughout the network. We also use the number of parameter metric to be able to compare models in terms of memory consumption to save the trained model. This metric was calculated by summing the number of weights needed throughout the model.

Table 4 Rank-1 accuracy and complexity (M: number of parameters and T: GFLOPS) of baseline and pruned ResNet CNNs on ImageNet dataset [77]

Networks	rank-1	Performance measures	
		T	M
ResNet50	76.01	6.32	23.48
ResNet34	73.27	6.67	21.28
ResNet18	69.64	3.09	11.12
L1	71.85	2.96	11.90
Taylor	71.65	3.21	12.09
Auto-Balanced	71.97	2.96	11.90
Entropy	71.46	2.96	11.90
PSFP	71.57	2.96	11.90

5 Experimental results and discussion

5.1 Pruning on pretraining data

Table 4 shows the performance of baseline and pruned ResNet CNNs (backbone ResNet50 and smaller ResNet 18 and ResNet34 networks) on the ImageNet dataset. Results in this table provide an indication of the benefits of pruning only on a large source domain pretraining dataset. There are few variations in results among the pruning techniques. The similarity in the results indicates that efficient pruning techniques rely on the availability of the large-scale datasets. Given the large-scale dataset in this experimental evaluation, ranking done by each technique is very similar even though the ranking metrics differ. Pruning results with ResNet50 very similar to the state of the art without pruning, although half the FLOPS are required. For example, using ThiNet [71] provides a compressed network that yields rank-1 accuracy of around 72% for half the FLOPS. Results with pruned networks provide higher accuracy than the smaller ResNet18 while having similar computational complexity and memory consumption.

For person ReID datasets, we attempt to preserve the same pruning compression ratio of 50% for comparison. This ratio is the highest compression level while minimizing the difference in the results between the baseline and the pruned models. We also prune the same number of filters per layer, use the same number of pruning iteration, and same fine-tuning iteration, layer-by-layer. Across layers is not the same, 50% filter gone, 5% filter at iteration, and stopping condition is 50% pruned away.

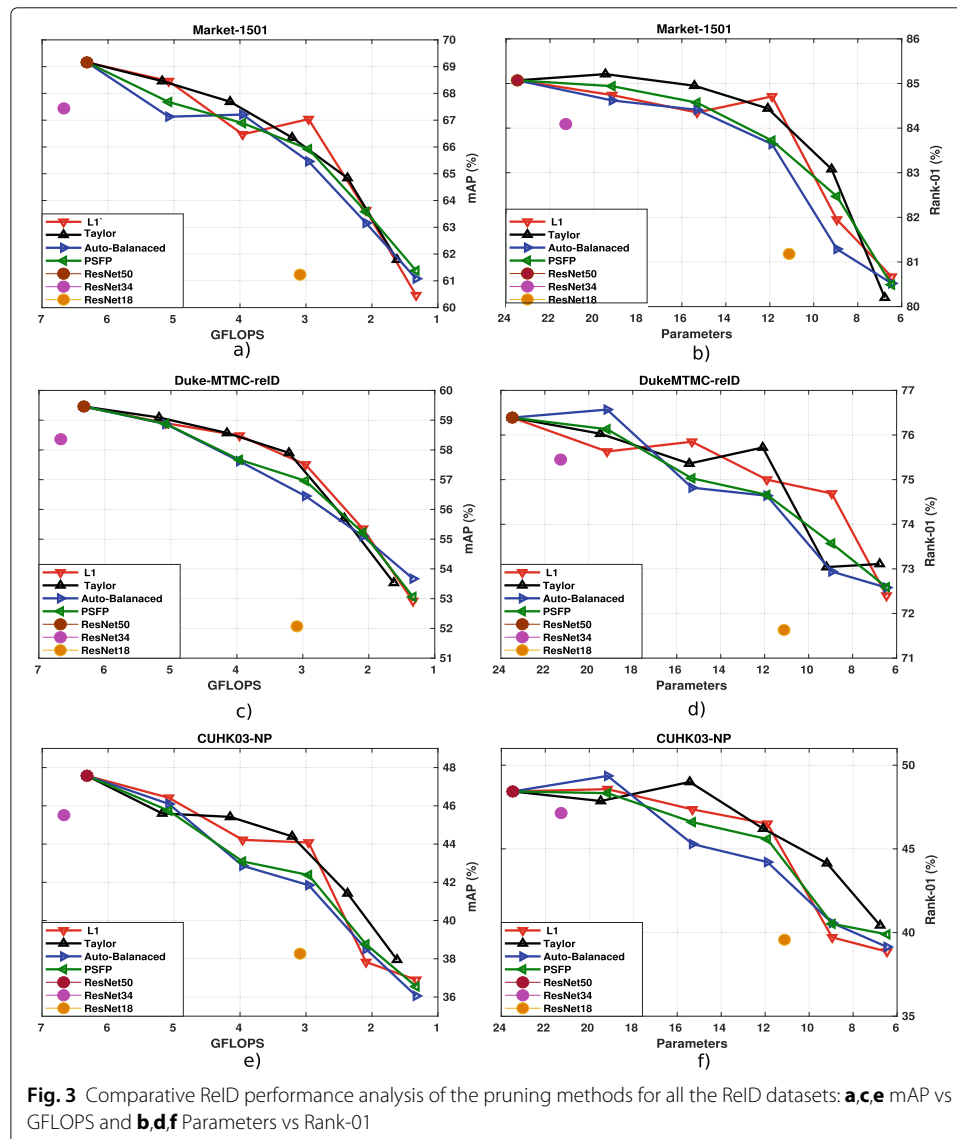
Table 5 reports the results for Market-1501, DukeMTMC-reID, and CUHK03-NP ReIDs. The reported results are for all the scenario. Taylor has higher FLOPS and a higher number of parameters than the other methods which would probably lead to a slower model and more consumption in terms of memory. Out of the 5 methods, the L1 method seems to be working the best by having the best or close to the best on the three datasets.

The pruned models also have shown less performance drop in terms of accuracy while reducing considerably the number of FLOPS and parameters. Pruned models are faster than backbone ResNet50 network while having similar performance (around 1%). Plus, the pruned models have a similar number of FLOPS and parameters to ResNet18 while having better results on the three performance metrics. This means that pruning a larger model is more advantageous than using a shallower model like ResNet18.

To get a more global view of these results, the graphics in Fig. 3 depicts visually which models are better where the optimal placement would be top right and the worse would

Table 5 Accuracy (mAP and rank-1:R-1) and complexity (M: parameters and T: GFLOPS) of baseline and pruning Siamese networks on ReID datasets

Networks	M	T	Market-1501		DukeMTMC		CUHK03-NP	
			mAP	rank-1	mAP	rank-1	mAP	rank-1
ResNet50	23.48	6.32	69.16	85.07	59.46	76.39	47.57	48.43
ResNet34	21.28	6.67	67.44	84.09	58.36	75.45	45.51	47.14
ResNet18	11.12	3.09	61.23	81.18	52.07	71.63	38.27	39.57
L1	11.90	2.96	67.04	84.71	57.51	75.00	44.08	46.50
Taylor	12.09	3.21	66.35	84.44	57.90	75.72	44.40	46.21
Auto-Balanced	11.90	2.96	65.46	83.64	56.45	74.64	41.85	44.21
Entropy	11.90	2.96	65.16	82.39	56.64	74.64	42.44	44.07
PSFP	11.90	2.96	65.92	83.72	56.96	74.66	42.38	45.58



be bottom left. There are two graphics for each dataset where the first one presents the mAP vs FLOPS and the second one presents Rank1 vs Parameters.

5.2 Pruning on target application data with weak ReID baseline

The objective of this experiment is to analyze and compare the pruning techniques with weak ReID baseline such as *Trinet* [3]. Table 6 reports the experimental evaluations of all the scenarios. For fair comparison, we chose to keep the compression ratio to 5% of the total number of channels. Our Scenario 2 results are produced using the HaoLi Iteration 3 model as the model pruned on the pretraining dataset. Using the same model for the considered techniques gives us a better idea on which pruning technique is the best when we pruned directly on our task dataset.

As we can observe in Table 6, the results with pruning directly on the target operational domain are not performing as good as the performances of the same pruned model with large-scale pretraining dataset. We can make the following observations from these results: (1) pruning and fine-tuning should be done on the same domain as in the case of Scenario 1 and Scenario 3, no matter whether it is source or target operational domain; (2) lack of data in target domain affects the pruning accuracy to regain the information loss by the pruning of the weak channels; (3) with the large-scale source dataset and the L1 method, we were able to prune our model to the same number of FLOPS as Scenario 2 (2.09 GFLOPS) but our Rank1 accuracy was 81.95% instead of 70.67%. The L1 method also seems to be better suited to prune directly on the task dataset compared to Taylor and Entropy. This might be explained by the fact that we do not have many samples per person since Taylor and Entropy approach uses a subset of samples to determine which channels to prune compared to the L1 method that ranks the channels with their weights; (4) Scenario 4 is not viable since all methods' performance drop drastically. (5) As for the Auto-Balanced and the PSFP techniques, they seem to outperform the other methods. This could be explained by the fact that auto-balanced modifies the loss in order to transfer the information of the pruned channels to the remaining one. This scheme seems to help considerably when our number of samples is limited. The PSFP seems to be the best suited algorithm to pruned models on a limited dataset. This can probably be explained by the fact that we only zeroed the pruned channels which keep the model architecture which allows the recovery of certain soft-pruned channels during the fine-tuning phase. Our results with the PSFP are also very similar to the ones obtained with the Scenario 1 scheme where we prune our models on the large-scale source dataset and then fine-tune on our task domain dataset. The great advantage of this method is the fact that we can prune and fine-tune our models in the same step. Plus, we are skipping the slow step of pruning on the very large ImageNet dataset.

To compare the scenarios further, we used two compression ratios which are around half the FLOPS (C1) and around one-third (C2) of the FLOPS of the original ResNet50 model. The Scenario 2 model for the first compression is using the second iteration L1 model as the model pruned on ImageNet. As for the second compression, we are using the third iteration. The results for the following experiments are found in Table 7.

Table 7 shows us that Scenario 1 is truly the best one since all the results outperform the other ones for any method and any dataset. As for the comparison between Scenario 2 and 3, the conclusion to determine which one is better is hard to make

Table 7 Comparison of network accuracy (mAP and rank-1) and complexity (M: memory (parameters) and time: T (GFLOPS)) with different pruning compression ratios of different scenarios on all the ReID datasets

Datasets	Scenario (compression)	M	T	L1		Auto-Balanced		PSFP	
				mAP	rank-1	mAP	rank-1	mAP	rank-1
Market-1501	1 (C1)	11.90	2.96	67.04	84.71	65.46	83.64	65.92	83.72
	2 (C1)	11.90	2.96	47.46	69.36	47.57	70.46	65.55	82.69
	3 (C1)	11.90	2.96	53.22	72.21	54.73	74.05	65.88	82.19
	1 (C2)	8.95	2.09	63.63	81.95	63.16	81.29	63.58	82.47
	2 (C2)	8.95	2.09	49.18	70.67	47.36	68.74	65.03	80.85
	3 (C2)	8.95	2.09	41.93	62.62	48.30	69.00	65.88	82.91
Duke-MTMC	1 (C1)	11.90	2.96	57.51	75.00	56.64	74.64	56.96	74.66
	2 (C1)	11.90	2.96	40.60	59.47	41.09	61.09	53.71	71.90
	3 (C1)	11.90	2.96	46.88	65.93	45.54	66.16	56.62	74.09
	1 (C2)	8.95	2.09	55.35	74.69	55.10	72.94	55.22	73.57
	2 (C2)	8.95	2.09	40.91	61.67	43.19	64.14	53.05	73.20
	3 (C2)	8.95	2.09	39.17	58.71	34.80	54.58	56.77	73.38
CUHK03-NP	1 (C1)	11.90	2.96	44.08	46.50	41.85	44.21	42.38	45.58
	2 (C1)	11.90	2.96	27.23	28.43	27.44	29.57	40.47	45.00
	3 (C1)	11.90	2.96	33.57	36.07	33.34	35.29	40.66	44.57
	1 (C2)	8.95	2.09	37.83	39.71	38.51	40.57	38.76	40.52
	2 (C2)	8.95	2.09	26.29	27.36	26.69	27.36	42.19	43.86
	3 (C2)	8.95	2.09	26.79	28.29	26.50	27.14	40.31	40.14

since Scenario 2 can be done using many configurations to get to a model similar to the one in Scenario 3. We could either prune more on the large-scale source dataset or prune less. Scenario 2 results are also affected by the choice of the pruned model on the large-scale source set. Our first compression results using the second iteration of the L1 method perform less in terms of recognition accuracy than pruning only on the target operational dataset (Scenario 3). But using the third iteration as shown in the second compression results, our Scenario 2 results are better than our Scenario 3 results.

5.3 Pruning on target application data with strong ReID baseline

The results shown in Table 7 indicate that PSFP so far is the best performing pruning approach in most scenarios. Additionally, since PSFP is suitable for deploying a compressed model—training can be done while the pruning is applied—we apply this technique on a strong ReID baseline. Therefore, the aim of this experiment is to analyze the effectiveness of the pruning techniques using a strong ReID baseline PCB [43]. We show two experimental analyses with PSFP pruning of PCB architectures. The first experiment shows the effect of ReID accuracy while pruning only backbone feature extractor (indicated as PCB (BFE)), while the second one considers all the layers (i.e., local convolutional layers and fully connected layers) after backbone architectures those perhaps use for feature compression and for classification tasks. In addition to the original ResNet [11], we also performed experiments with SE-ResNet [78] to see the effectiveness of pruning methods on different backbone CNNs.

Table 8 Comparison of network accuracy (mAP and rank-1) and complexity (M: Parameters and T: GFLOPS) with different pruning scenarios on all the person ReID datasets. BFE backbone feature extractor; LC, local convolutional layer; FC, fully connected layer

Methods	Backbone	mAP	Market-1501			DukeMTMC-reID			
			rank-1	T	M	mAP	rank-1	T	M
PCB (Baseline)		77.3	92.4	6.1	27.2	65.3	81.9	6.1	27.2
PSFP + PCB (BFE)	ResNet50	69.4	90.4	2.6	12.0	60.7	78.5	2.6	12.0
PSFP + PCB (BFE+ LC + FC)		69.1	89.1	2.6	11.16	59.4	77.8	2.6	11.16
PCB (Baseline)		77.3	91.7	5.9	27.7	65.3	81.2	5.9	27.7
PSFP + PCB (BFE)	SE-ResNet50	70.0	88.4	2.59	14.5	61.9	78.9	2.59	14.5
PSFP + PCB (BFE+ LC + FC)		69.2	87.2	2.58	13.69	59.4	77.4	2.58	13.69

Experimental results for the strong baseline PCB are reported in Table 8 both for Market-1501 and DukeMTMC-reID datasets. Our results show a consistency with the initial claims—the number of FLOPS and parameters required by PCB’s ResNet and SE-ResNet architectures are reduced by half, while maintaining a comparable rank-1 accuracy for both ReID datasets. Results also suggest that PSFP pruning of local convolutional layers and FC layers have little effect on ReID accuracy as the margin of differences between *PSFP+PCB(BFE)* and *PSFP+PCB(BFE+LC+FC)* is small. This analysis implies that it is worth pruning backbone architecture rather with local convolutional and FC layers since it allows more memory and parameter reduction. It is worth noting that the margin of decline in mAP accuracy is higher than that of rank-1 accuracy for both backbones, and on all ReID datasets.

5.3.1 Filter selection criteria

As a part of ablation study, this experiment aims to analyze the effect of magnitude-based filter selection criteria such as l_p -norm on ReID accuracy. We conducted this experiment with PSFP+PCB(BFE) on ReNet50. We show a comparative ReID performance analysis between l_1 -norm and l_2 -norm on Table 9. It can be observed from Table 9 that the ReID performance of l_2 -norm criteria is marginally better than that of l_1 -norm criteria. This is due to the effect of the largest element that has been dominant in l_2 -norm. As a consequence, the filters with largest weights preserved while pruning provide more discriminative features for better recognition accuracy.

5.3.2 Varying pruning rates

The objective of this experiment is to observe ReID performance when varying pruning rates. It was performed with PSFP+PCB(BFE). Figure 4a and b show the mAP and rank-01 accuracy obtained with varying pruning rates, respectively. With both measures, the

Table 9 Comparison of network accuracy (mAP and rank-1) and complexity (M: Parameters and T: GFLOPS) with different pruning criteria of the PSFP approach on all the person ReID datasets. BFE, backbone feature extractor

Methods	Criteria	mAP	Market-1501			DukeMTMC-reID			
			rank-1	T	M	mAP	rank-1	T	M
PSFP+PCB(BFE)	l_1 -norm	68.7	90.0	2.6	12.0	60.6	78.8	2.6	12.0
PSFP+PCB(BFE)	l_2 -norm	69.4	90.4	2.6	12.0	60.7	78.5	2.6	12.0

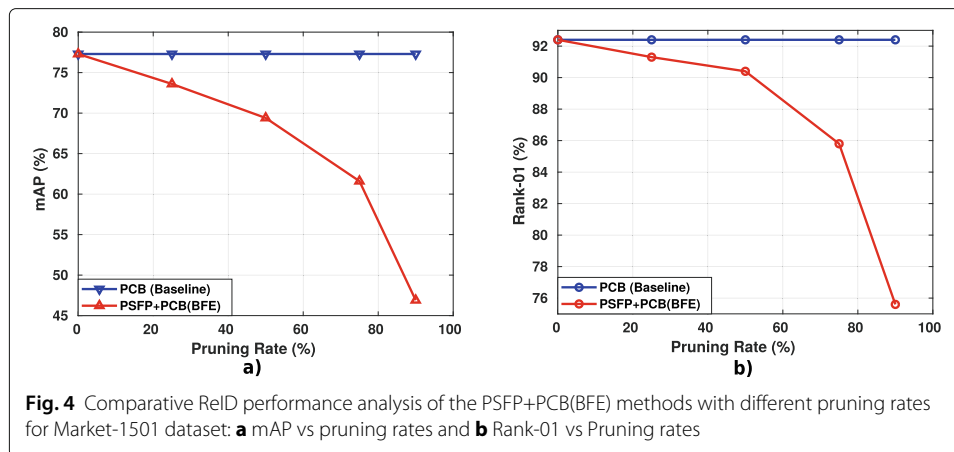


Fig. 4 Comparative ReID performance analysis of the PSFP+PCB(BFE) methods with different pruning rates for Market-1501 dataset: **a** mAP vs pruning rates and **b** Rank-01 vs Pruning rates

accuracy of the pruned model drops exponentially with growing pruning rates. For pruning rates between 0 and 25%, the accuracy of the pruned model drops marginally. The pruning rate above 50% leads to drastic decline in ReID performance. When pruning a larger number of filters, the loss of information affects accuracy considerably.

To further analyze the pruning on target operational domain, we apply the best fine-tuning practices proposed in [73], as presented in Section 3.4. We have calculated their metrics for ImageNet and Market11501 and got 0.005 for the cosine distance and 2.45 for MMD. With these metrics and the fact that Market1501 has fewer than 20 samples for each class, the authors proposed to freeze the feature extractor during the fine-tuning to avoid overfitting since our task dataset is small and close to our large-scale pretraining dataset. Since our problem of a small dataset was showing during the retraining phase of the pruned network, we decided to try prune one layer at a time and freeze the others during the retraining phase. The goal of this strategy is to force the pruned layers to relearn the loss information while maintaining the other layers in the same optimal region as the baseline model. This method was tried for Scenario 2 with the L1 method. We decided to prune layer 5 of the ResNet50 while freezing the rest of the network. The model was pruned to 2.61 GFLOPS and the rank1 accuracy was 76.10%. This experiment shows that we could limit the effects of pruning by using a layer-by-layer approach and freezing the other layers to regain the accuracy. The problem with this scheme is that it is not very effective time-wise since it is a long and fastidious task to prune and retrain to the desired compression ratio for each layer instead of doing the whole model in one pass.

6 Conclusions

In this paper, we exploit the prunability of the state-of-the-art pruning models that are suitable for compressing deep architecture for person ReID application in terms of criteria to select channels and of strategies to reduce channels. In addition to that, we propose different scenarios or pipelines for leveraging a pruning method during the deployment of a network for a target application. Experimental evaluations on multiple benchmarks source and target datasets indicate that pruning can considerably reduce network complexity (number of FLOPS and parameters) while maintaining a high level of accuracy. It also suggests that pruning larger CNNs can also provide a significantly better performance than fine-tuning smaller ones. One key observation of the scenario-based

experimental evaluations is that pruning and fine-tuning should be performed in the same domain.

Future experiments could explore a reduction in pruning iterations in order to reduce the impact of pruning on knowledge corruption. Retraining of the pruned networks could also be improved by adding a learning rate decay. Using layer-by-layer methods, with different compression ratios for each layer, can improve the results since some layers are more resilient to pruning than others. Techniques for freezing parts of the network can also improve accuracy, but drastically increase the time complexity for pruning and retraining phases. The soft pruning method could also benefit from better selection criteria, e.g., using a gradient-based approach instead of the norm of the channel weights. Finally, another interesting future experiment would be to avoid costly pruning on large pretraining dataset and only use the progressive soft pruning scheme to see if it can achieve similar results with higher compression ratios. While pruning approaches have proven to be effective in person ReID, we realized that it is focused on the same domain, which can limit its usage. In addition, work on pruning in the unsupervised learning settings is still quite limited. Future work could extend such pruning methods to unsupervised domain adaptation in person ReID.

Acknowledgements

The authors would like to thank all the state-of-the-art works for sharing their code which help us to validate our approach by setting their approaches as baselines.

Authors' contributions

All the authors contributed by participating in the discussion, experimentation, and drafting the manuscript of the work described in this paper. All authors read and approved the final manuscript.

Funding

This work was supported by the Mitacs Accelerate Master's Fellowship, Elevate Postdoctoral Fellowship Program, and the Natural Sciences and Engineering Research Council of Canada.

Availability of data and materials

The re-identification datasets used to validate the findings of this work are publicly available that can be downloaded following the given references of each dataset.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 26 September 2020 Accepted: 3 June 2021

Published online: 25 June 2021

References

1. J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, K. Murphy, in *Proc. IEEE Conf. Vis. Pattern Recognit. (CVPR)*, Speed/accuracy trade-offs for modern convolutional object detectors, (Honolulu, 2017), pp. 7310–7311
2. E. Ahmed, M. Jones, T. K. Marks, in *Proc. IEEE Conf. Vis. Pattern Recognit. (CVPR)*, An improved deep learning architecture for person re-identification, (Boston, 2015), pp. 3908–3916
3. A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
4. R. R. Viorio, M. Haloi, G. Wang, in *Proc. European Conf. Comput. Vis. (ECCV)*, Gated siamese convolutional neural network architecture for human reidentification (Springer, Amsterdam, 2016), pp. 791–808
5. W. Chen, X. Chen, J. Zhang, K. Huang, in *Proc. IEEE Conf. Vis. Pattern Recognit. (CVPR)*, Beyond triplet loss: a deep quadruplet network for person reidentification, (Honolulu, 2017), pp. 403–412
6. M. Geng, Y. Wang, Y. Shi, K. Yan, M. Geng, Y. Tian, T. Xiang, in *Proc. IEEE Conf. on Multimedia Big Data (BigMM)*, Deep transfer learning for person reidentification, (Xi'an, 2018), pp. 1–5
7. D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, in *Proc. IEEE Conf. Vis. Pattern Recognit. (CVPR)*, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, (Las Vegas, 2016), pp. 1335–1344
8. H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, End-to-end comparative attention networks for person re-identification. *IEEE Trans. on TIP.* **26**(7), 3492–3506 (2017)
9. A. Bhuiyan, Y. Liu, P. Siva, M. Javan, I. B. Ayed, E. Granger, in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision*, Pose guided gated fusion for person reidentification, (Aspen, 2020), pp. 2675–2684

10. A. Bhuiyan, A. Perina, V. Murino, in *Proc. European Conf. Comput. Vis. (ECCV)*, Person re-identification by discriminatively selecting parts and features (Springer, Zurich, 2016), pp. 147–161
11. K. He, X. Zhang, S. Ren, J. Sun, in *Proc. IEEE Conf. Vis. Pattern Recognit. (CVPR)*, Deep residual learning for image recognition, (Las Vegas, 2016), pp. 770–778
12. S. Han, H. Mao, W. J. Dally, in *International Conference on Learning Representations (ICLR)*, Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding, (San Juan, 2016). Conference Track Proceedings
13. S. Han, J. Pool, J. Tran, W. Dally, in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), vol. 1*, Learning both weights and connections for efficient neural network, (Montreal, 2015), pp. 1135–1143
14. Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, R. Feris, in *Proc. IEEE Conf. Vis. Pattern Recognit. (CVPR)*, Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification, (Honolulu, 2017), pp. 5334–5343
15. R. Rigamonti, A. Sironi, V. Lepetit, P. Fua, in *Proc. IEEE Conf. Vis. Pattern Recognit. (CVPR)*, Learning separable filters, (Portland, 2013), pp. 2754–2761
16. E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, R. Fergus, in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), vol. 1*, Exploiting linear structure within convolutional networks for efficient evaluation, (Montreal, 2015), pp. 1269–1277
17. M. Jaderberg, A. Vedaldi, A. Zisserman, in *Proceedings of the British Machine Vision Conference*, Speeding up Convolutional Neural Networks with Low Rank Expansions (BMVA Press, Nottingham, 2014)
18. V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets, V. Lempitsky, in *3rd International Conference on Learning Representations, ICLR*, Speeding-up convolutional neural networks using fine-tuned CP-decomposition, (San Diego, 2015). Conference Track Proceedings 2015
19. C. Tai, T. Xiao, Y. Zhang, X. Wang, E. Weinan, in *4th International Conference on Learning Representations, ICLR*, Convolutional neural networks with low-rank regularization, (San Juan, 2016). Conference Track Proceedings 2016
20. T. Cohen, M. Welling, in *International conference on machine learning (ICML)*, Group equivariant convolutional networks (PMLR, New York City, 2016), pp. 2990–2999
21. H. Van Hasselt, A. Guez, D. Silver, in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30, No. 1*, Deep reinforcement learning with double q-learning, (Phoenix, 2016)
22. W. Shang, K. Sohn, D. Almeida, H. Lee, in *international conference on machine learning (ICML)*, Understanding and improving convolutional neural networks via concatenated rectified linear units (PMLR, New York City, 2016), pp. 2217–2225
23. S. Dieleman, J. De Fauw, K. Kavukcuoglu, in *International conference on machine learning (ICML)*, Exploiting cyclic symmetry in convolutional neural networks (PMLR, New York City, 2016), pp. 1889–1898
24. C. Bucilua, R. Caruana, A. Niculescu-Mizil, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Model compression, (2006), pp. 535–541
25. G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
26. P. Luo, Z. Zhu, Z. Liu, X. Wang, X. Tang, et al, in *AAAI*, Face model compression by distilling knowledge from neurons, (2016), pp. 3560–3566
27. T. Chen, I. Goodfellow, J. Shlens, Net2net: accelerating learning via knowledge transfer. arXiv preprint arXiv:1511.05641 (2015)
28. Y. Gong, L. Liu, M. Yang, L. Bourdev, in *6th International Conference on Learning Representations, ICLR*, Compressing deep convolutional networks using vector quantization, (Vancouver, 2018). Conference Track Proceedings 2018
29. W. Chen, J. Wilson, S. Tyree, K. Weinberger, Y. Chen, in *International conference on machine learning (ICML)*, Compressing neural networks with the hashing trick (PMLR, Lille, 2015), pp. 2285–2294
30. Y. Le Cun, J. S. Denker, S. A. Solla, in *Proceedings of the 2nd International Conference on Neural Information Processing Systems (NIPS)*, Optimal brain damage (MIT Press, Cambridge, 1989), pp. 598–605
31. B. Hassibi, D. G. Stork, in *Proceedings of the 5th International Conference on Neural Information Processing Systems (NIPS)*, Second order derivatives for network pruning: optimal brain surgeon, (Denver, 1992), pp. 164–171
32. P. Molchanov, S. Tyree, T. Karras, T. Aila, J. Kautz, Pruning Convolutional Neural Networks for Resource Efficient Transfer Learning. CoRR abs/1611.06440 (2016)
33. H. Li, A. Kadav, I. Durdanovic, H. Samet, H. P. Graf, Pruning Filters for Efficient ConvNets. CoRR abs/1608.08710 (2016)
34. J.-H. Luo, J. Wu, An Entropy-based Pruning Method for CNN Compression. CoRR abs/1706.05791 (2017)
35. Y. He, X. Zhang, J. Sun, in *Proceedings of the IEEE International Conference on Computer Vision*, Channel pruning for accelerating very deep neural networks, (2017), pp. 1389–1397
36. Y. He, X. Dong, G. Kang, Y. Fu, Y. Yang, in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, Progressive deep neural networks acceleration via soft filter pruning, (Stockholm, 2018), pp. 2234–2240
37. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, in *Proceedings of the IEEE international conference on computer vision*, Scalable person re-identification: A benchmark, (Santiago, 2015), pp. 1116–1124
38. W. Li, R. Zhao, T. Xiao, X. Wang, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Deepreid: Deep filter pairing neural network for person re-identification, (Columbus, 2014), pp. 152–159
39. E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, in *Proc. European Conf. Comput. Vis. (ECCV)*, Performance measures and a data set for multi-target, multi-camera tracking (Springer, Amsterdam, 2016), pp. 17–35
40. D. Yi, Z. Lei, S. Liao, S. Z. Li, in *IEEE 22nd International Conference on Pattern Recognition*, Deep metric learning for person re-identification, (Columbus, 2014), pp. 34–39
41. A. Hermans, L. Beyer, B. Leibe, In Defense of the Triplet Loss for Person Re-Identification. arXiv e-prints arXiv-1703 (2017)
42. L. Zheng, Y. Yang, A. G. Hauptmann, Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
43. Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, in *Proceedings of the European conference on computer vision (ECCV)*, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), (Munich, 2018), pp. 480–496

44. C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, in *Proceedings of the IEEE international conference on computer vision*, Pose-driven deep convolutional model for person reidentification, (Venice, 2017), pp. 3960–3969
45. H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, Q. Tian, Deep representation learning with part loss for person re-identification. *IEEE Trans. Image Process.* **28**(6), 2860–2871 (2019)
46. L. Zhao, X. Li, Y. Zhuang, J. Wang, in *Proceedings of the IEEE international conference on computer vision*, (CVPR), Deeply-learned part-aligned representations for person re-identification, (Honolulu, 2017), pp. 3219–3228
47. S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification. *Pattern Recog.* **48**(10), 2993–3003 (2015)
48. R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, X. Bai, in *Proceedings of the European conference on computer vision (ECCV)*, Hard-aware point-to-set deep metric for person reidentification, (Munich, 2018), pp. 188–204
49. N. Wojke, A. Bewley, in *2018 IEEE winter conference on applications of computer vision (WACV)*, Deep cosine metric learning for person re-identification, (Lake Tahoe, 2018), pp. 748–756
50. L. Wu, C. Shen, An entropy-based pruning method for cnn compression. arXiv preprint arXiv:1706.05791 (2017)
51. J. Bromley, I. Guyon, Y. LeCun, E. Säcker, R. Shah, in *Advances in neural information processing systems (NIPS)*, vol. 6, Signature verification using a “Siamese” time delay neural network, (Denver, 1992), pp. 737–44
52. R. R. Viorar, B. Shuai, J. Lu, D. Xu, G. Wang, in *European Conference on Computer Vision*, A siamese long short-term memory architecture for human re-identification (Springer, Amsterdam, 2016), pp. 135–153
53. F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Joint learning of single-image and cross-image representations for person re-identification, (Las Vegas, 2016), pp. 1288–1296
54. K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Omni-scale feature learning for person re-identification, (Seoul, 2019), pp. 3702–3712
55. H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Bag of tricks and a strong baseline for deep person reidentification, (Seoul, 2019)
56. Z. Dai, M. Chen, X. Gu, S. Zhu, P. Tan, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Batch dropout network for person re-identification and beyond, (Seoul, 2019), pp. 3691–3701
57. Y. Shen, T. Xiao, H. Li, S. Yi, X. Wang, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, End-to-end deep kronecker-product matching for person reidentification, (Salt Lake City, 2018), pp. 6886–6895
58. N. Martinel, G. L. Foresti, C. Micheloni, Deep pyramidal pooling with attention for person re-identification. *IEEE Trans. Image Process.* **29**, 7306–7316 (2020)
59. M. Zheng, S. Karanam, Z. Wu, R. J. Radke, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Re-identification with consistent attentive siamese networks, (Long Beach, 2019), pp. 5735–5744
60. G. Chen, C. Lin, L. Ren, J. Lu, J. Zhou, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Self-critical attention learning for person re-identification, (Seoul, 2019), pp. 9637–9646
61. A. Wu, W.-S. Zheng, X. Guo, J.-H. Lai, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Distilled person re-identification: towards a more scalable system, (Long Beach, 2019), pp. 1187–1196
62. F. Hafner, A. Bhuiyan, J. F. Kooij, E. Granger, A cross-modal distillation network for person re-identification in rgb-depth. arXiv preprint arXiv:1810.11641 (2018)
63. Z. Liu, J. Qin, A. Li, Y. Wang, L. Van Gool, in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, Adversarial binary coding for efficient person re-identification, (Shanghai, 2019), pp. 700–705
64. D. Mekhazni, A. Bhuiyan, G. Ekladios, E. Granger, in *European Conference on Computer Vision, Online*, Unsupervised domain adaptation in the dissimilarity space for person re-identification (Springer, 2020), pp. 159–174
65. W. Fang, H.-M. Hu, Z. Hu, S. Liao, B. Li, Perceptual hash-based feature description for person re-identification. *Neurocomputing.* **272**, 520–531 (2018)
66. S. Gong, J. Cheng, Z. Hou, et al., in *European Conference on Computer Vision, Online*, Faster person re-identification (Springer, 2020), pp. 275–292
67. B. O. Ayinde, J. M. Zurada, Building efficient convnets using redundant feature pruning. arXiv, 2018 (2018)
68. Y. He, P. Liu, Z. Wang, Y. Yang, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Filter pruning via geometric median for deep convolutional neural networks acceleration, (Long Beach, 2019)
69. P. Singh, V. K. Verma, P. Rai, V. P. Namboodiri, in *28th International Joint Conference on Artificial Intelligence (IJCAI)*, Play and Prune: Adaptive filter pruning for deep model compression, (Macao, 2019), pp. 3460–3466
70. X. Ding, G. Ding, J. Han, S. Tang, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, No. 1, Auto-balanced filter pruning for efficient convolutional neural networks, (New Orleans, 2018)
71. J.-H. Luo, H. Zhang, H.-Y. Zhou, C.-W. Xie, J. Wu, W. Lin, Thinet: pruning CNN filters for a thinner net. *TPAMI*, 2018 (2018)
72. R. Yu, A. Li, C.-F. Chen, J.-H. Lai, V. I. Morariu, X. Han, M. Gao, C.-Y. Lin, L. S. Davis, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Nisp: Pruning networks using neuron importance score propagation, (Salt Lake City, 2018), pp. 9194–9203
73. B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, T. Darrell, Best practices for fine-tuning visual classifiers to new domains. *Proc. European Conf. Comput. Vis. (ECCV)*, 435–442 (2016)
74. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**(3), 211–252 (2015)
75. Z. Zhong, L. Zheng, D. Cao, S. Li, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Re-ranking person re-identification with k-reciprocal encoding, (Honolulu, 2017), pp. 1318–1327
76. Z. Zheng, L. Zheng, Y. Yang, in *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, (Honolulu, 2017), pp. 3754–3762
77. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, in *2009 IEEE conference on computer vision and pattern recognition (CVPR)*, Imagenet: A large-scale hierarchical image database, (Miami, 2009), pp. 248–255
78. J. Hu, L. Shen, G. Sun, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Squeeze-and-excitation networks, (Salt Lake City, 2018), pp. 7132–7141

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.