Medium Article Link: https://medium.com/@adityatushar.wadnerkar/pruning-methods-for-person-re-identification-a-survey-860cf8a987cf

SlideShare Presentation Link: https://www.slideshare.net/AdityaWadnerkar1/pruning-methods-for-person-reidentification-a-survey

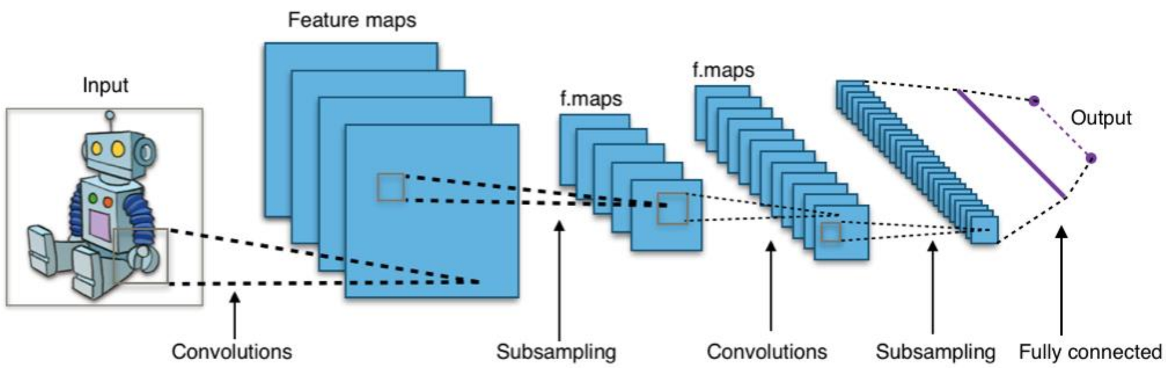
Pruning Methods for Person Re-identification: A Survey

Original Author:

Hugo Massona, Amran Bhuiyana, Le Thanh Nguyen-Meidinea, Mehrsan Javanb, Parthipan Sivab, Ismail Ben Ayeda, Eric Grangera
Article Link: https://arxiv.org/abs/1907.02547
Date Published: July 04 2019

Introduction

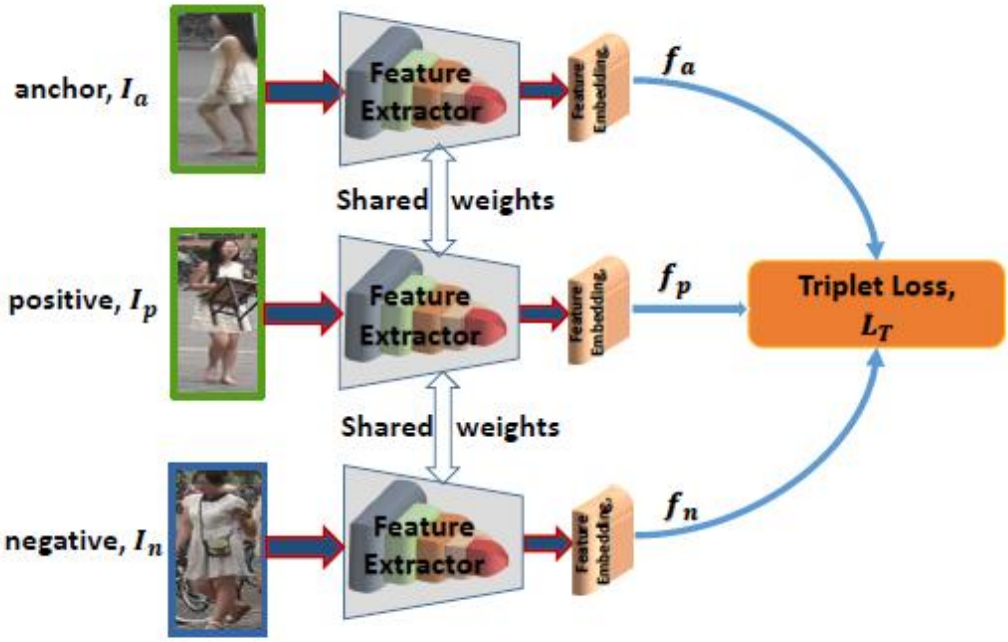
We have seen tremendous development in the field of deep learning in every sector of the industry with applications ranging from speech recognition to self-driving cars. One such development is in its application for recognizing people. This is one of the applications which has a wide scope not only for the industry, but also in the daily lives of people. There are numerous methods which are used for re-identifying people. However, most methods require optimization of the algorithms for training and testing. There are various pruning techniques that help us do and, in this article, we will discuss about them in detail.



What are CNNs?

Convolutional Neural Networks or ConvNets are a type of deep neural networks widely used for image processing. These contain an input, output and multiple hidden layers and perform convolution, a special kind of linear mathematical operation, instead of using regular matrix multiplication in more than one of its layers. However, as the complexity of tasks increases complexity of CNNs increases due to wider and deeper networks.

Introduction to Siamese Networks



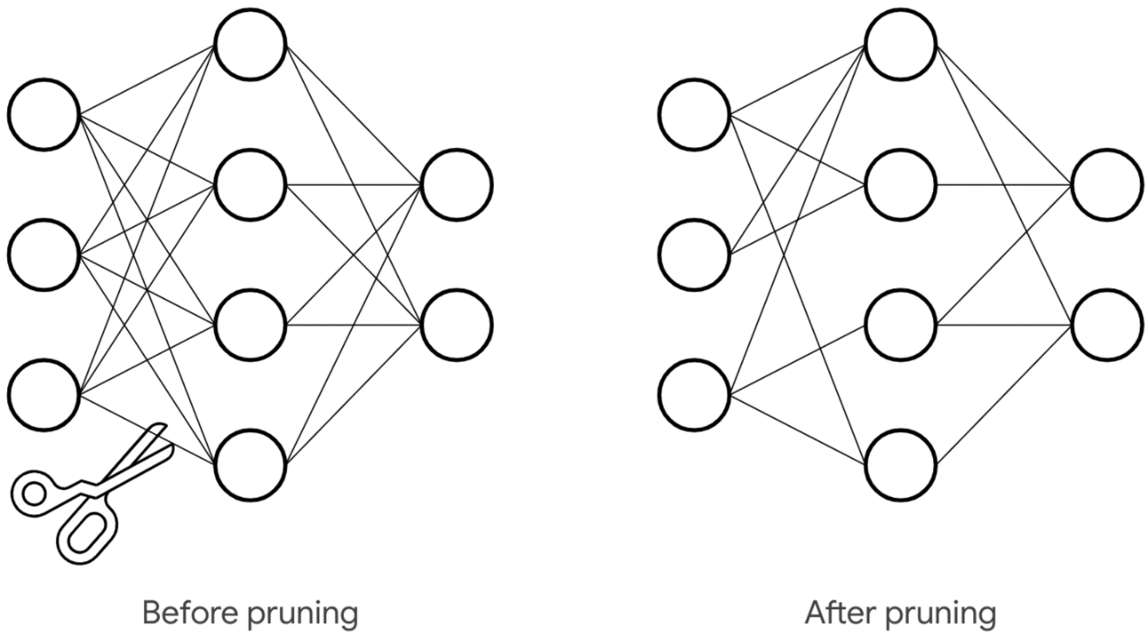
Siamese neural networks are also called as twin neural networks and belong to the class artificial neural network. These work parallelly on two different input vectors and compute output vectors which are comparable. One of the most popular application of Siamese networks is face recognition, which involves comparison with pretrained images of people. Most person reidentification methods use pretrained CNN due to their exceptional performance.

$$\mathcal{L}_{\mathrm{T}} = \frac{1}{N_T} \sum_{\substack{\mathrm{a,p,n} \\ y_a = y_p \neq y_n}} \left[m + d\left(\mathrm{f}_a, \mathrm{f}_p\right) - d\left(\mathrm{f}_a, \mathrm{f}_n\right) \right]_+$$

$$\mathcal{L}_{\mathrm{TBH}} = \frac{1}{N_s} \sum_{a=1}^{N_s} \left[m + \max_{y_p = y_a} d\left(\mathrm{f}_a, \mathrm{f}_p\right) - \min_{y_p \neq y_a} d\left(\mathrm{f}_a, \mathrm{f}_n\right) \right]_+$$

We sample a triplet of images Ia, Ip and In for a mini-batch, {Ia,Ip} and {Ia,In} being pair of image for same and different individuals respectively and the features for backbone network being fa, fp and fn. We initially sample triplet for a person and then sample pairs and compute loss in following steps. At last we compare most positive and negative value obtained from computations.

Pruning Techniques for CNN



Before we dive deep into the techniques, let's discuss what pruning of algorithms exactly means. Pruning is used to reduce algorithm complexity by removing non-essential (or less essential) parameters from neural networks. Removing unnecessary features won't affect much of the accuracy while reducing the complexity and computational requirements.

Table 1: A Taxonomy of techniques according to pruning strategyto reduce chanels.

Pruning Strategy	Methods
Prune Once	Hao Li[31] Redundant Channels[40] Entropy[32]
Iterative Pruning	Molchanov[30] Play and Prune[41] FPGM[42]
Pruning using regularization	Auto-Balance[43] Play and Prune[41]
Pruning by minimizing reconstruction error	ThiNet[44] Channel Pruning[33]
Progressive Pruning	PSFP[34]

While pruning neural networks we need to take care of a few things: the pruning criteria which needs to be able to identify the factors which are major contributors to the accuracy as compared to those which are not, the appropriate ratio for compression as there is a trade-off between reducing complexity and losing accuracy and at last, the scheduling of retraining and pruning in multiple iterations as doing everything in one iteration may cause significant damage.

Pruning Schemes

In PSFP pruning scheme the model doesn't lose its original dimension during training phase and as per proposed changes, adding a progressive pruning with increased compression ratios can lead to a shallower network.

Algorithm 1 Algorithm Description of PSFP

Input: training data: $\mathbf{X}$
Input: pruning rate: P_i, pruning rate decay D
Input: the model with parameters $\mathbf{M} = \{\mathbf{M}^{(i)}, 0 \leq i \leq L\}$
  Initialize the model parameter $\mathbf{M}$
  for $epoch = 1; epoch \leq epoch_{max}; epoch++$ **do**
    Update the model paramters $\mathbf{M}$ based on X
    for $i = 1; i \leq L; i++$ **do**
      Calculate the l_2-norm for each channel
      Calculate the pruning rate P' at this epoch using P_i and D
      Select the N lowest l_2-norm depending on the pruning rate
      Zeroize the weights W of the selected channels
    end for
  end for
  Obtain the compact model with parameters **M'** from $\mathbf{M}$
Output: Compact model with parameters M'

It is proposed to add a progressive pruning scheme where at each pruning iteration, the compression ratio is increased in order to get a shallower network. After completion of pruning the channels with lowest ranking are discarded based on their compression ratio.

FPGM prunes using geometric median to prune channels. It is also denoted as Play and Prune technique which doesn't focus on criterion, rather finds the ideal number of channels which can be pruned at given error tolerance rate.

Algorithm 2 Algorithm Description of FPGM

Input: training data: $\mathbf{X}$
Input: pruning rate: P
Input: the model with parameters $\mathbf{M} = \{\mathbf{M}^{(i)}, 0 \leq i \leq L\}$
  Initialize the model parameter $\mathbf{M}$
  for $epoch = 1; epoch \leq epoch_{max}; epoch++$ **do**
    Update the model parameters $\mathbf{M}$ based on X
    for $i = 1; i \leq L; i++$ **do**
      Select the $n_{out} \times P$ of W_i channels that satisfy Equation 22
      Zeroize the selected channels
    end for
  end for
  Obtain the compact model with parameters **M'** from $\mathbf{M}$
Output: Compact model with parameters M'

Table 3: Comparison of rank-1 accuracy and network complexity analysis in term of GFLOPS and Parameters taken from the literature.

Dataset	CIFAR10					
Feature Extractor	ResNet56					
Algorithm	Original			Pruned		
	R-1 (%)	GFLOPS	Parameters (M)	R-1 (%)	FLOPS (G)	Parameters (M)
Hao Li [31]	93.04	0.125	0.85	93.06	0.091	0.73
Auto-Balanced [43]	93.93	0.142	N/D	92.94	0.055	N/D
Redundant channel [40]	93.39	0.125	0.85	93.12	0.091	0.65
PP [41]	93.39	0.125	0.85	93.09	0.039	N/D
FPGM [42]	93.39	0.125	0.85	92.73	0.059	N/D

Dataset	ImageNet					
Feature Extractor	VGG16					
Algorithm	Original			Pruned		
	R-1 (%)	GFLOPS	Parameters (M)	R1 (%)	GFLOPS	Parameters (M)
ThiNet [44]	90.01	30.94	138.34	89.41	9.58	131.44
Molchanov [30]	89.30	30.96	N/D	87	11.5	N/D
HaoLi [31]	90.01	30.94	138.34	89.13	9.58	130.87
Channel Pruning [33]	90.01	30.94	138.34	88.1	7.03	131.44

Dataset	ImageNet					
Feature Extractor	ResNet50					
Algorithm	Original			Pruned		
	R-1 (%)	GFLOPS	Parameters (M)	R-1 (%)	GFLOPS	Parameters (M)
Entropy [32]	72.88	3.86	25.56	70.84	2.52	17.38
ThiNet [44]	75.30	7.72	25.56	72.03	3.41	138
FPGM [42]	75.30	7.72	25.56	74.83	3.58	N/D

The trade-off between complexity of algorithm, simplifying steps of pruning with losing accuracy rate highly influences choice of criteria. In case the pruning and training requires fast deployment due to time issue, it depends on L1 and L2, else with no time constraints minimization can yield best performance at cost of high computation.

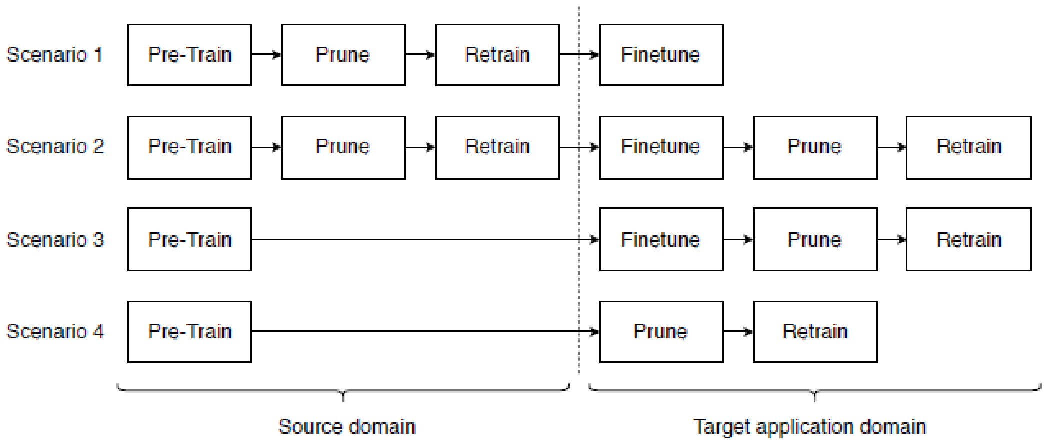


Figure 3: Scenarios for pruning and training a CNN.

Datasets

There are four datasets which are publicly available: ImageNet, Market1501, CUHK03-NP and DukeMTMC-reID. **ImageNet** is divided into two parts: The first part has is 1.2M images for training the model and the second part has 50k for validation. The dataset contains 1000 natural images classes. **Market-1501** contains 1501 entities captured by different cameras, and 32,668 pedestrian image bounding-boxes. **CUHK03-NP** has 14,096 images with 1,467 entities. It captures each person using two cameras with 4.8 images on average in each camera on the CUHK campus. **DukeMTMC-reID** has 1,812 entities and is constructed with multi-camera tracking dataset, with 702 identities used as training set and other 1,110 entities as testing set. ImageNet is used as pre-trained dataset and rest of datasets are used for testing person-identification.

Performance Analysis

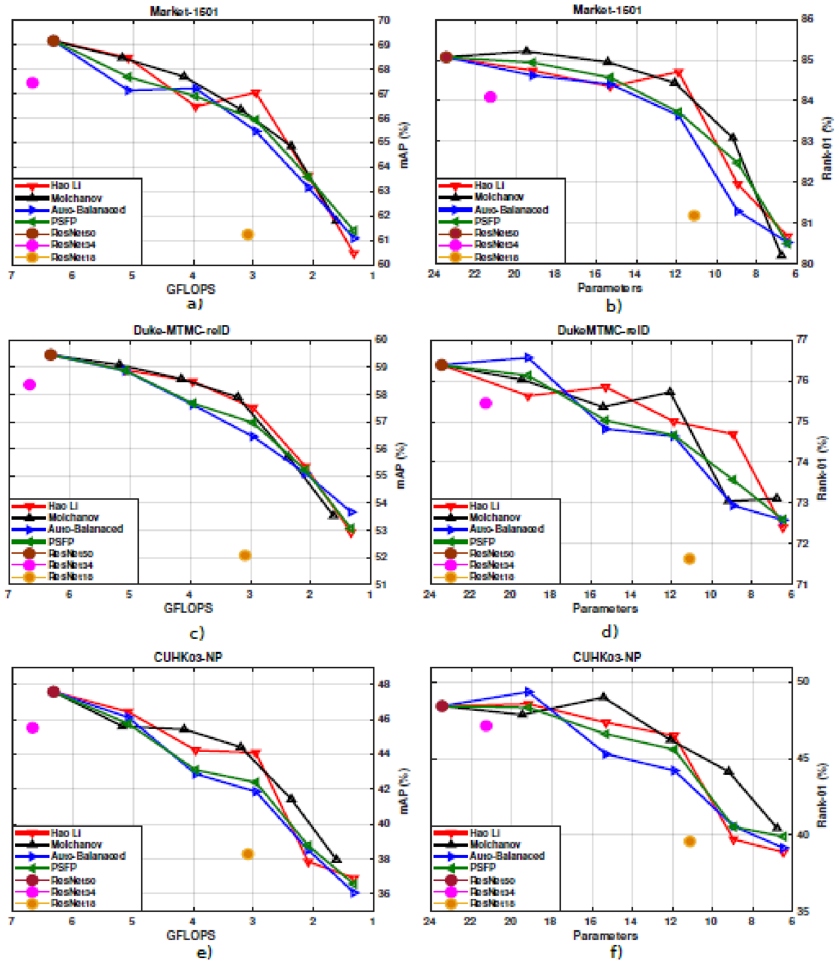
Table 5: Accuracy and complexity of baseline and pruning Siamese networks on ReID datasets. Mean average precision (mAP) and rank-01 accuracy (R-1) are shown in percentage (%).

Networks	Parameters	GFLOPS	Market-1501		DukeMTMC		CUHK03-NP	
			mAP	R-1	mAP	R-1	mAP	R-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
HaoLi	11.90	2.96	67.04	84.71	57.51	75.00	44.08	46.50
Molchanov	12.09	3.21	66.35	84.44	57.90	75.72	44.40	46.21
AutoBalanced	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

Comparing the 5 methods, the Hao Li method outperformed the rest by having the best results on the three datasets. Molchanov was found to be working slower and consuming more memory than the other methods due to higher number of FLOPS and parameters.

The figure lets is visualise performances of models to understand which models performed better than the rest with two graphics for all datasets; first presenting mAP vs FLOPS and second presenting Rank1 vs Parameters.

Approaching layer by layer and freezing rest of layers can help retain accuracy, but problem arises when it’s not very effective with time. The reason being that pruning layer by layer and retaining accuracy is difficult task as compared to doing the same with just one pass.



Conclusion

We talked about various pruning techniques for person re-identification by compressing Siamese networks, criteria selecting channels, and methods to reduce channels. We have also discussed different pipelines integrating pruning methods for application during deployment of network. Experimental evaluations on multiple benchmarks source and target datasets indicate that pruning can considerably reduce network complexity, i.e. number of FLOPS and parameters, while maintaining a high level of accuracy. A key observation from the scenario based experimental evaluations is that both fine tuning and pruning should be performed in the same domain. Moreover, pruning larger CNNs can also provide a significantly better performance than fine tuning the smaller ones.

#Deep Learning #Convolutional #Neural Networks #Siamese Networks #Complexity #Pruning #Domain #Adaptation #Person Re-identification

Original Paper - https://arxiv.org/abs/1907.02547