

**PROPOSAL AND IMPLEMENTATION OF
AN EFFECTIVE CNN BASELINE FOR PERSON RE-
IDENTIFICATION**

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF
MASTER OF TECHNOLOGY
IN
INFORMATION SYSTEMS

Submitted By:

SUMIT KUMAR

(2K19/ISY/16)

Under the Supervision

of

DR. DINESH KUMAR VISHWAKARMA

(Professor Department of IT)



DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY

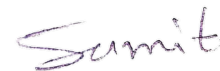
October 2021

CANDIDATE'S DECLARATION

I Sumit Kumar, Roll No. 2K19/ISY/16 student of M. Tech Information Systems, hereby declare that the project report titled “**Proposal and implementation of an effective CNN baseline for person re-identification**” which is submitted by me to the Department of Information Technology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

October 26, 2021



Sumit Kumar

CERTIFICATE

I hereby certify that the Project Report titled “**Proposal and implementation of an effective CNN baseline for person re-identification**” which is submitted by Sumit Kumar, Roll No 2K19/ISY/16, Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

October 26, 2021



Dr. Dinesh Kumar Vishwakarma

SUPERVISOR

ACKNOWLEDGMENT

I express my gratitude to my major project guide **Dr. Dinesh Kumar Vishwakarma**, Professor Department of IT, Delhi Technological University, for the valuable support and guidance he provided in making this major project. It is my pleasure to record my sincere thanks to my respected guide for his constructive criticism and insight without which the project would not have shaped as it has.

I humbly extend my words of gratitude to other faculty members of this department for providing their valuable help and time whenever it was required.



Sumit Kumar

Roll No. 2K19/ISY/16

M. Tech (Information System)

ABSTRACT

Person re-identification is a challenging task due to the critical issues of human pose variation, human body occlusion, camera view variation, etc. To deal with this, most of the state-of-the-art methods based on the deep convolutional neural networks have strong feature extraction and classification capacity. However, there are not enough studies about building an effective CNN baseline model. There are three good practices are followed in this work for building an effective CNN architecture. These practices are adding batch normalization after the global pooling layer, use only one fully connected layer for classification and use Adam optimizer. Using these three techniques in the implementation, the performance of a simple pre-trained CNN model have been enhanced without making any high level changes and experimental results supports this argument

TABLE OF CONTENTS

TITLE	PAGE NO.
CANDIDATE'S DECLARATION	ii
CERTIFICATE	iii
ACKNOWLEDGMENT	iv
ABSTRACT	v
Figures and Tables	vii
1. INTRODUCTION	1
1.1 Person re-identification	1
1.2 Applications	2
1.3 Difficulties	2
1.4 Diversification in problem	3
1.5 Thesis outline	3
2. BACKGROUND	5
2.1. Overview	5
2.2. Deep Learning	5
2.3. Person reID approaches	11
2.4. Datasets and evaluation metrics	14
3. PROPOSED MODEL	16
4. EXPERIMENT	19
4.1 Experimental setup	19
4.2 Results	22
5. CONCLUSION	25

5.1 Challenges	25
5.2 Future work	26
REFERENCES	27

LIST OF TABLES

TABLE	TITLE	PAGE NO.
Table 1:	Specifications of some widely used datasets for person re-identification	14
Table 2:	Performance evaluation and key issues of some generally used datasets in Person re-ID	15
Table 3:	Performance evaluation on Market-1501 dataset	22
Table 5:	Performance on DukeMTMC-reID dataset	23

LIST OF FIGURES

FIGURE	TITLE	PAGE NO.
Figure 1.1:	Matching a person over different frames	1
Figure 2.1:	Biological Neuron and Artificial Neuron	5
Figure 2.2	Activation functions	6
Figure 2.3:	Feedforward neural network and recurrent neural network	7
Figure 2.4:	Convolution operation	8
Figure 2.5:	Pooling operation	8
Figure 2.6:	Convolutional Neural Network	9
Figure 2.7:	LeNet-5 architecture	10
Figure 2.8:	Residual learning block used in Deep Residual networks	10
Figure 2.9:	Descriptors	12
Figure 3.1:	Pipeline of proposed CNN based approach	16
Figure 4.1:	Image samples from datasets	19
Figure 4.2:	Training and validation loss curve and top1 error curve	21
Figure 4.3:	Training and validation top1 accuracy curve	22

CHAPTER 1

INTRODUCTION

1.1 Person Re-identification

Person re-identification is recognizing any person after being already observed by different camera views. Usually, a person appears in disparately in different cameras, therefore person re-identification is challenging. Poor quality of CCTV images makes it more challenging to identify a person only by facial recognition that's why we have to add more feature like walking pattern, color, texture, clothing etc. along with facial recognition in order to increase performance. Features like color and texture are commonly used in person re-identification because these features have better generalization capability.

There are several state of the arts methods of person re-identification that achieved extraordinary performance on different datasets, but still there are some datasets for which achieving good performance by a person re-identification method is a challenging. We can understand this variation in performance of different datasets by discussing their different features and limitations.

Person re-identification gained lots of attention from computer vision society in the current years especially after the introduction of deep learning methods in person re-identification problems. Several methods of deep learning have been introduced to achieve this task. Deep CNNs (convolutional neural networks) introduced lots of changes and significant progress is achieved. The main advantage of using deep convolutional neural networks is that they can perform feature extraction, classification, and metric learning together in an optimistic way.



Fig 1.1 Matching a person over different frames

1.2 Applications

Person re-identification has vast range of experiential applications from surveillance to medical applications. The major practical applications of person re-identification are:

- **Tracking people in cross-camera:** In surveillance systems, for understanding the whole scenario it requires to track persons over multiple cameras, it is helpful in analysing the crowd or any activities in crowded places. Person re-identification is used to track the movement of a person by establishing correspondence between tracks.
- **Image retrieval:** In this case, query image of target person is searched in a large database and according to similarity ranked list is provided.
- **Human-Robot interaction:** In case of robot-human interaction, identity of the dialogist is sustained.
- **Analysing human behaviour or activity for long term:** This application is useful in retail and health care. For example, customers shopping behaviour can be analysed by observing that which products customers is seeing or touching and in healthcare it can be used to assist doctors.

1.3 Difficulties

This is very challenging for a machine to recognize a person over different camera views because of various reasons like poor image quality, variance in appearance, hidden face, same texture etc. Some main challenges in person re-id are discussed below: -

1. **Person detection:** - At first, in order to re-identifying any person, person has to be detected and bounding boxes must be defined for person in an image.
2. **Poor image quality:** - Some CCTV cameras have low resolution and this deficiency of information makes person re-identification a more difficult problem.
3. **Person Re-Id in a video:** - when we are dealing with video input then the process of establishing unison amidst subject detected across frames is called **tracking**. Tracking multiple individuals is challenging task.
4. **Obstacle:** - Feature extraction is difficult due to partial or complete occlusion of person in crowded places.
5. **Variation in Illumination:** - Same subject can appear in different shades or colors in different camera views due to variation in the sharpness of sunlight, shade, reflection of light from reflective surfaces, indoor lighting etc.

6. **Identical clothing:** - Extraction of information from clothing is difficult at places like schools or some workplaces where people have uniform clothing.
7. **Need good number of images for per id training:** As any person can appear for very less times in a network of cameras, so it becomes difficult to gather significant amount of data for single person due to which it becomes difficult to train a good model.
8. Camera position, labeling of data, limitations in real time scenario.

1.4 Diversification in problem

Variation in input: - Person re-identification can be performed on both image and video input. In case of video input gait analysis can be used.

Person re-identification can be mainly classified in two types on the basis of whether query already present in gallery or not. When query's id already in gallery then it is called open set re-identification. Example- when camera is in open environment then new person's features are extracted. On the other hand, when ID of query is possibly present in gallery then query is matched with individual's ID with highest similarity.

1.5 Thesis Outline

In the next chapter, a brief discussion of literature is presented. Additionally, an overview of existing datasets is also provided. Apart from the literature review the main objective is to propose an effective CNN baseline for person re-identification.

In this thesis, we mainly focused on building some general rules for making an effective CNN baseline without using any high-level approaches. Summary of following chapters is:

In Chapter 2 we will discuss about the deep learning techniques for person re identification and various state of the art techniques are discussed. Some CNN models which are pre-trained on the ImageNet are also discussed in this chapter. After introducing deep learning and techniques for person re-identification we will discuss about some existing datasets for person re-identification. At the end a brief introduction of evaluation matrices is also given.

In Chapter 3 we will present the proposed techniques for improving the performance of pre-trained CNN architectures for person re-identification. Here, there is discussion about three techniques or rule that are followed for building CNN baseline model using pre-trained

network for person re-identification and will theoretically prove the argument that using these three practices will improve the performance and will reduce the over-fitting issue.

In Chapter 4 we will experimental results which support the argument given in chapter 3. First we will talk about experimental setup, datasets and evaluation metrics. And finally, comparisons between implemented approach and some state of the art methods will we shown.

Chapter 5 is about the future work and the conclusion of the thesis.

CHAPTER 2

BACKGROUND

2.1 Overview

In this chapter, we have two sections, first section discusses about deep learning in which we will start with introducing neural networks and after that we will discuss about convolution neural network (CNN), some of its variants and pre-trained CNNs.

2.2 Deep Learning

In current times, deep learning has become the powerful and popular subset of machine learning after introduction of the neural networks. Deep learning algorithms use multiple processing layers and have shown high performance. Most of the deep learning models have outperformed the traditional state of the art approaches in the field of computer vision and NLP (natural language processing). Let's discuss the main concepts of deep learning:

2.2.1 Neural Network

As we can guess from the term “neural”, it is inspired from the functioning biological neurons of human brain. Human behavior is determined by nervous system which contains neurons. Human brain contains about 85 billion neurons; these neurons react with each other and transfer the information.

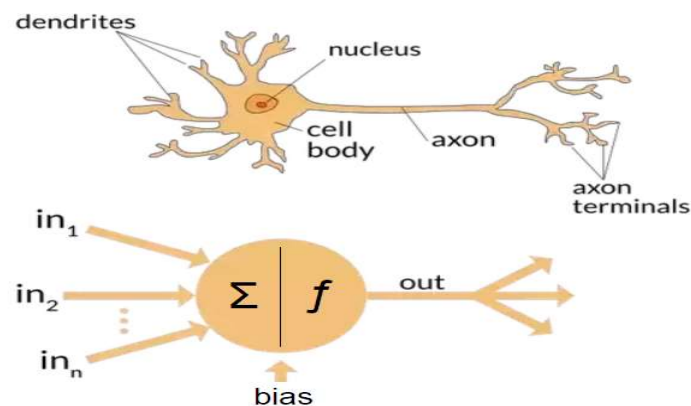


Fig. 2.1 Biological Neuron and Artificial Neuron

An artificial neuron is mathematically modeled as a function which takes input vector x and calculates weighted sum with weight vector w and adds bias b and transform the result using activation function α which is nonlinear:

$$y = \sigma(\sum_i w_i x_i + b) \tag{2.1}$$

There are many activation functions. Initially purposed activation functions were based on thresh holding but used very less due to its un-derivability. Currently, mainly used activations are the logistic function, tanh, sigmoid function or ReLU (Rectified Linear Unit) function.

The curve of the sigmoid function looks like an S-shaped curve. Sigmoid function is utilized since it exists between 0 to 1. Hence, for the models where we must predict the probability as an output it is particularly useful. Since the value of probability can only be in between 0 and 1, we can say that sigmoid is a good option. Tanh is like the arranged sigmoid and is a better option and it is also S-shaped curve. The range of tanh function is ranges between -1 to 1. In this graph, the benefit is that Tanh graph maps the negative and zero inputs. At present, the most widely utilized activation function is ReLU. Since then, it's been utilized in nearly all CNNs and deep learning systems.

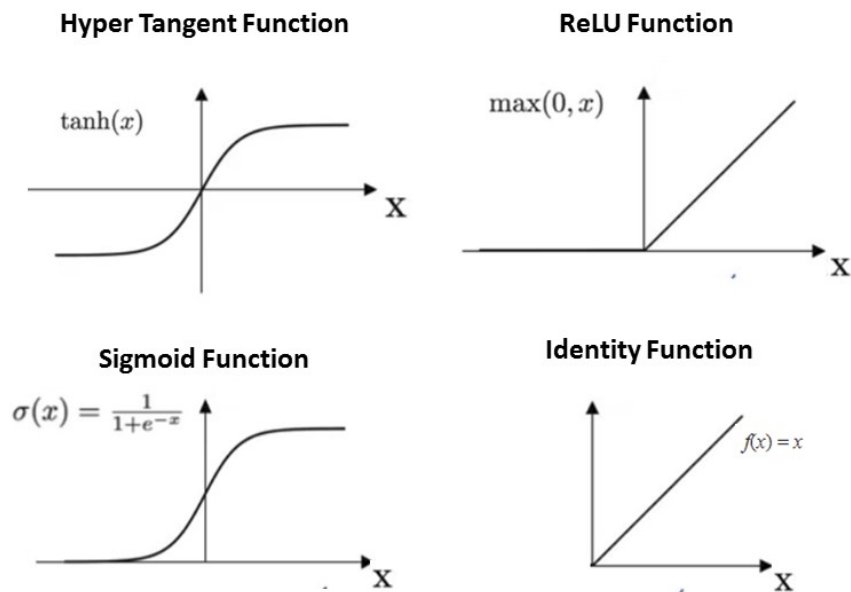


Fig. 2.2 Activation functions

Since, an artificial neuron can only simply calculate the weighted sum, it cannot work as a good classifier. We should use a network of artificial neurons in order to deal with complex

tasks. Theoretically neurons can be arranged randomly but in practical applications they are ordered in way that a neuron's input cannot depend upon its output. Neural networks in which activated outputs propagate forward are called feed-forward neural networks but recurrent neural networks can contain cycles which makes it more efficient for making dynamic models but due to cycles training becomes difficult.

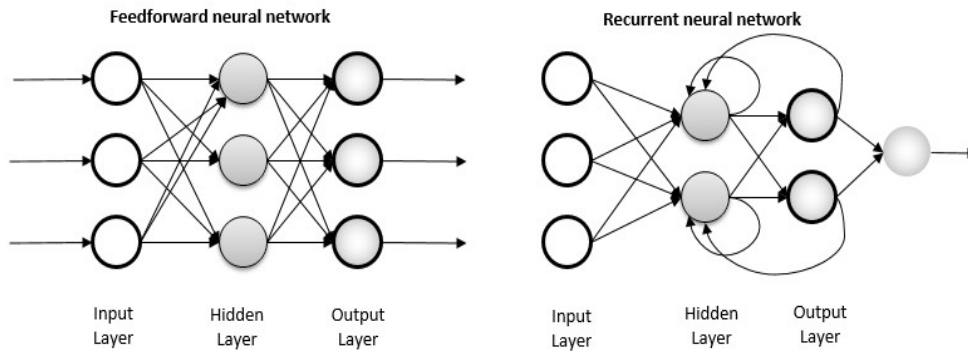


Fig. 2.3 Feedforward neural network and recurrent neural network

2.2.2 Convolutional Neural Network

In computer vision tasks we deal with images and images can have many hundreds to tens of thousands of pixels in it and if we process this much input data only with the help of fully connected neural networks then it requires very large number of parameters to train the model there is high possibility that it may lead to over-fitting. To overcome with this problem convolutional layer were introduced. For reducing the complexity convolution layers use shared weights. A convolutional neural network gets input in matrix form and convolutional layers use weight matrix which is also called kernel or filter which is applied on input matrix. For a given two-dimensional input image X when kernel K is applied then convolution operation can be mathematically defined according to equation 2.2 as:

$$(W * X)(i, j) = \sum_m \sum_n X(m, n)W(i - m, j - n) \quad (2.2)$$

After applying kernel on input matrix, the resultant matrix is called feature map. There are different kernels for different applications for example edge detection, blurring, sharpening, loopy pattern detection etc. The values in these kernels are different depending upon their applications. The coefficients of convolutional kernel matrix are calculated automatically by using back propagation algorithm. There are many kernels and feature maps in one convolution Layer.

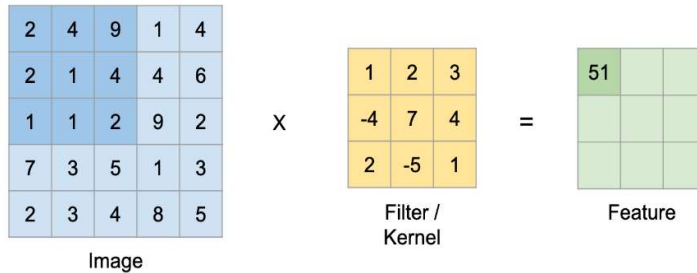


Fig. 2.4 Convolution operation

A convolutional neural network contains convolutions layers, pooling layers and fully connected layers. We already discussed about convolution layer and its convolution operation. In a CNN there can be many convolution layers and after each convolutional layer there is pooling layers. Pooling layers are used for dimensionality reduction, it does so by using a pooling filter over the output matrix of previous convolution layer, there are three types of pooling operations: average pooling, min pooling and max pooling. In average pooling, all values under applied kernel are replaced with average of the batch while in min and max pooling, minimum and maximum values in the batch are selected.

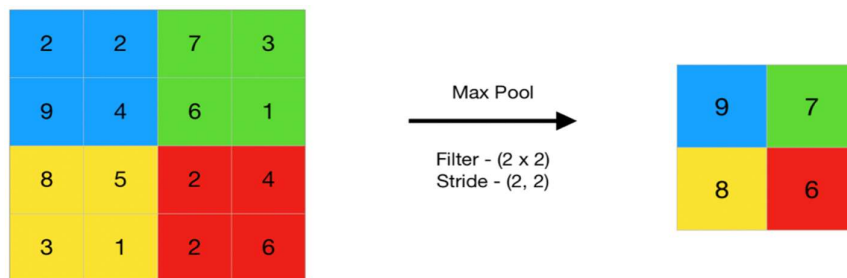


Fig. 2.5 Pooling operation

Among all pooling operations max pooling the most commonly used. Max pooling helps in position invariant feature detection. Below are the advantages of pooling:

- Reduces dimensions and computation
- Reduces over-fitting as there is less number of parameters after pooling
- Makes the model tolerant towards small distortions and variations.

Convolution neural network works in two phases, first is feature extraction phase which contains many convolutions and pooling layers and second part is classification which uses fully connected layers. CNNs mostly deal with two-dimensional inputs and for classification we need one-dimensional input, so after the last pooling layer of the feature extraction phase we have to flatten the output of the last pooling layer and then feed it to the fully connected layer.

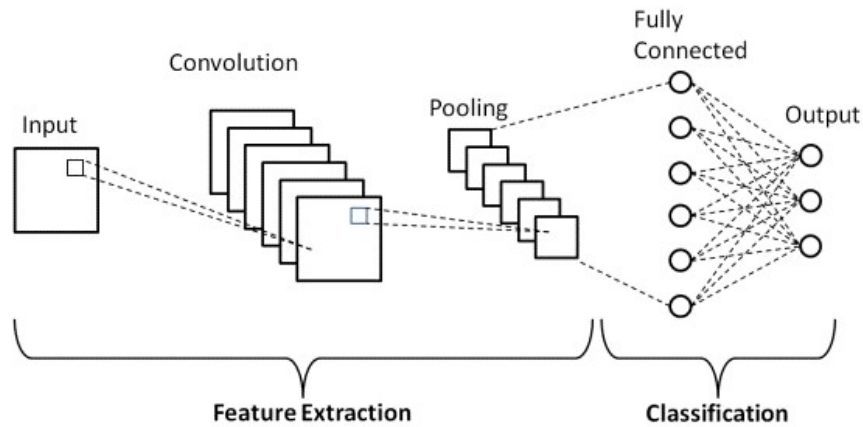


Fig. 2.6 Convolutional Neural Network

2.2.3 CNN Architectures

In this section, we will discuss about various widely used CNN architectures. It will help us to understand how basic CNN was improved to achieve state-of-the-art performance.

LeNet [1] is a simple classic CNN model developed for handwritten digit recognition in the early 90s. This architecture has three types of layers: convolution operation is performed by a convolution layer and features are automatically learned by backpropagation of the pooling layer. A pooling layer is used to deliver a summarized result to the next hidden layer, and at the last, there is a fully connected layer to perform the classification.

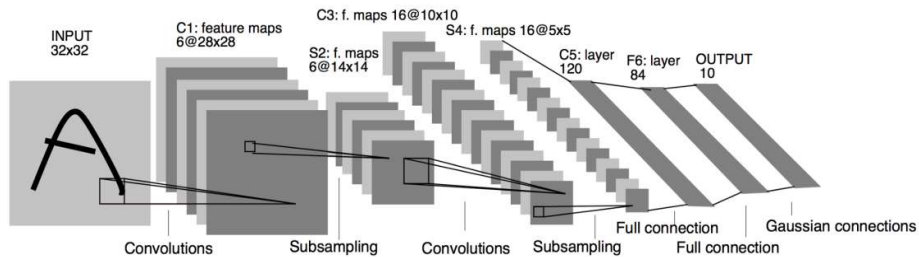


Fig.2.7 LeNet-5 architecture

AlexNet [2] is trained on ImageNet [3] dataset and it is also winner of ILSVRC 2012. ImageNet is a huge dataset which contains over 15,000,000 labeled images of around 22000 categories. In ILSVRC a subset of ImageNet having around thousand images in each 1000 categories is used.

AlexNet has 5 convolutional layers and after convolutional layer there are 3 fully connected layers. For nonlinear transformation rectified-linear-unit (ReLU) activation function is used. ReLU deals with gradient vanishing problem and computing it is also easy. Dropout layers are used after every fully connected layer to deal with over fitting, in which some units are randomly dropped.

VGGNet [4] Alex net's huge filters (eleven in the first and five in the second convolutional layers) were replaced with numerous filters size of 3×3 one after the other by VGGNet.

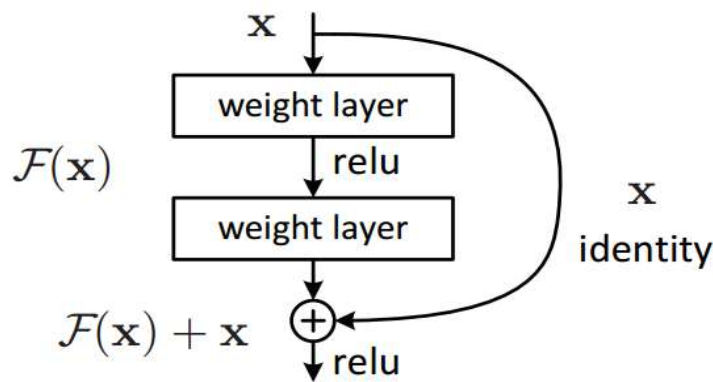


Fig. 2.8 Residual learning block used in Deep Residual networks

ResNet [5] (Residual Neural Network): Concept of residual block is introduced to solve the problem of vanishing gradient.

ResNet-50 architecture has 4 stages. Initial convolution kernel size is 7×7 , initial max-pooling kernel size is 3×3 .

- Stage-1: In stage-1 there are 9 layers in 3 residual blocks (3 layers within each) and size of kernels of these 3 layers is 64, 64 and 256.
- Stage-2: In stage-2 there are 12 layers in 4 residual blocks and size of kernels of these 3 layers is 128, 128 and 512.
- Stage-3: In stage-3 there are 18 layers in 6 residual blocks, size of kernels of these 3 layers is 256, 256 and 1024.

- Stage-4: In stage-4 there are 9 layers in 3 residual blocks size of kernels of these 3 layers is 512, 512 and 2048.
- After these four stages there is a fully connected layer of size 1000 and an average pooling layer.

2.3 Person Re-Id Approaches

A person re-identification algorithm mainly have two parts, first is appearance descriptor for representing a person and second is a matching function for comparing these descriptors. Throughout the years, lot of researches contributed to improve both the descriptor and matching algorithm. Recently, many deep learning approaches are introduced for person re-identification. Plenty of these methods have achieved the state of the arts performance by collectively learning the visual feature representation and similarity metric. In this section, person re-identification approaches and some commonly used datasets will be introduced.

2.3.1 Feature extraction methods

Similar to object recognition, we can use three features: shape, color and texture from static images to characterize the appearance of pedestrians. Widely used method for classifying color distribution is color histograms. Some normalization methods are proposed to make it robust for light intensity variations. Bak et al [6] used the ideal that uniformly distributed histogram has maximum entropy. Using color features only is not efficient when we are dealing with large gallery set because color of cloths can be similar in some people. Hence, shape and texture features are also often combined with color features. Shape features are based on edge or silhouette detection and these are subject to discrepancy in pose and viewpoint so that can be so important in identifying but are generally used in segmentation. Feature based on texture are the more effective. For getting a feature descriptor which robust and unique, many extraction strategies have been introduced which can be categorized into three categories: patch based, stripe based and body part based descriptors.

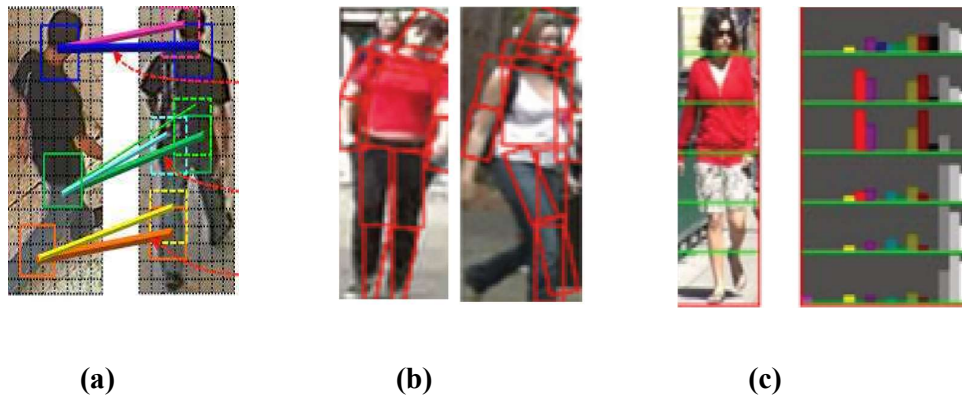


Fig.2.9 (a) Patch-based descriptor [7] (b) Body part based descriptor [8] (c) Stripe based descriptor [9]

2.3.2 Matching approaches

For matching based on extracted features, we can divide methods in two types. First is a matching function which learns in supervised way and second is distance metric learning in feature space.

Matching function learning

The idea is to calculate shortest possible distance between the descriptors for example, in histogram based descriptors Bhattacharya distance is used and in Euclidean space, L2 norms between the descriptors are used. Though, some features can be more important and can have more information for appearance matching. Many methods have been proposed for learning a matching function using a dataset of image pairs. Lin et al [10] had proposed a method which was based on learnt pair-wise dissimilarity profiles, where each person is treated as different class and problem is considered as a multi class classification problem. For solving person re-identification as a ranking problem, Prosser et al [11] proposed ensemble RanksSVM. In this method weak SVMs were trained on small datasets and then combined using boosting to make strong ranker.

Metric Learning: A metric which is specially learnt for images of persons is more discriminative and robust for variations in person images compared to generic distance measures like Euclidean or Bhattacharya distance.

2.3.3 Deep Learning based approaches

Recently, deep CNNs have achieved significant performance in the field of computer vision. In person re-identification, one of the applications of neural network is to learn embedding for which images are mapped to low dimensional feature space. One of the first works that used deep learning is done by Li et al. [12] that used Siamese neural nets to determine whether a pair of input images have the same identity. Most of the recent works on person re-identification mainly focus on creating baseline models based on deep convolutional neural networks. Pre-trained convolutional models trained on ImageNet are used by fine-tuning them on person re-identification datasets with the help of softmax loss. For image representation features from the last pooling layer are used. Recently, most of the deep learning based approaches adopting pre-trained models as backbone and trying to find other technical ways to further increase the performance of the re-identification system. Therefore, in most of the recent methods [13, 14] feature learned using only softmax loss usually serves as a baseline for comparison.

In person re-identification approaches many network architectures are being explored. ResNet50 [5] is the most commonly used architecture for backbone among person re-identification approaches but GoogleNet [15], DenseNet [16] and InceptionNet [17] are also being chosen by some researchers. By using pre-trained convolution models with metric learning the performance can be further improved.

When deep CNN models are trained on small datasets then the problem of over-fitting occurs. To deal with over-fitting many methods have been introduced. Especially for data augmentation, random erasing, random cropping and random flipping are commonly used techniques for training models. In recent techniques, dropout and batch normalization are widely used in training CNN models and are helpful in preventing over-fitting. In dropout technique, outputs of some random hidden neurons are discarded during training while batch normalization normalizes the output of each hidden neuron by using the mean and variance of the mini-batch.

2.4 Datasets and evaluation metrics

2.4.1 Datasets

There is a number of datasets that have been released for person re-identification. Some datasets for video-based person re-identification are also available. Some of the most commonly used benchmark datasets are summarized in table 1.

Table 1: Specifications of some widely used datasets for person re-identification [18]

Dataset	Year of Release	Number of individuals	Number of cameras	Number of images	Crop size
VIPeR	2007	632	2	1,264	128X48
iLIDS	2009	119	2	8,580	vary
ETH1,2,3	2007	85,35,28	1	8,580	Vary
GRID	2009	1,025	8	1,275	Vary
PRID2011z	2011	934	2	24,541	128X64
CUHK01z	2012	971	2	3,884	160X60
CUHK02z	2013	1,816	10	7,264	160X60
CUHK03	2014	1,467	10	13,164	Vary
CASIA Gait	2015	124	11	-	Vary
Market1501z	2015	1,501	6	32,217	128X64
PKU-Reid	2016	114	2	1,824	128X64
MARS	2016	1,261	6	1,19,1003	256X128
DukeMTMC-reID	2017	1,812	8	36,441	Vary
DukeMTMC4ReID	2017	1,852	8	46,261	Vary

These datasets have images from different environments. For example, images in iLIDS dataset are captured in airport arrival, GRID datasets contain images captured in underground station, CUHK01, CUHK02, CUHK03 and Market-150 have images from university campus. As we can see from Table 1, earlier, size of datasets was smaller in comparison to recent ones; scale of datasets is increased in recent years as they have over 1,000 IDs.

Table 2: Performance evaluation and key issues of some generally used datasets in Person re-ID [19]

Dataset	Key Issues	Highest accuracy (CMC Scores)			
		Rank 1	Rank 5	Rank 10	Rank 20
VIPeR	-	51.17	82.09	90.51	95.92
GRID	Image Quality is poor.	37.50	61.40	69.40	77.40
iLIDS	-	79.5	95.1	97.6	99.1
PRID2011	Some trajectories are not well-synchronized.	96.6	-	-	-

Dataset	Key Issues	Highest accuracy (CMC Scores)			
CUHK01	-	72.0	88.1	91.2	96.3
CUHK02	-	57.9	79.3	85.7	-
Market 1501	Quality of bounding boxes is poor.	98.0	98.9	99.1	-
DukeMTMC-reID	-	94.5	-	-	-

2.4.2 Evaluation Metrics

CMC (Cumulative matching characteristic) curve is mostly used metrics for evaluation of person re-identification algorithms. CMC curve shows probability that a query is present in different sized lists. In CMC calculation only first match is considered for calculation, so CMC is used as evaluation method when only one match is existing for each query. For a query, distance from query to each sample in gallery is calculated and rank is given to all gallery samples according to distance from query from small to large. Top-k accuracy is given by shifted step function. For final CMC curve computation, average of each shift step function is computed. Mean average precision (MAP) is also used for Market-1501 along with CMC because in Market-1501, multiple results exist from multiple cameras.

CHAPTER 3

PROPOSED METHOD

In this chapter, we will explain how to develop an effective baseline model of CNN for person re identification using three essential techniques. Because person re-identification datasets are relatively small, an effective method for preventing over-fitting is required for developing a high-accuracy person re identification model. Figure 1 depicts the major technical pipeline of proposed approach as well as three recommended best practices. The inputs are first fed into backbone network. The final subject representation is generated by feeding the "global pooling feature" provided by the last pooling layer into the batch normalization layer. We execute person identity categorization directly using only one fully-connected layer when using the batch-normalized feature. Removing multiple fully connected layers [14] is the procedure of dimension reduction. Finally, for training the CNN, Adam is used as the optimizer. These proposed techniques are simple, but it can be easily applied to a variety of CNN architectures. It is worth noting that in our approach, only softmax loss is used as a supervision signal.

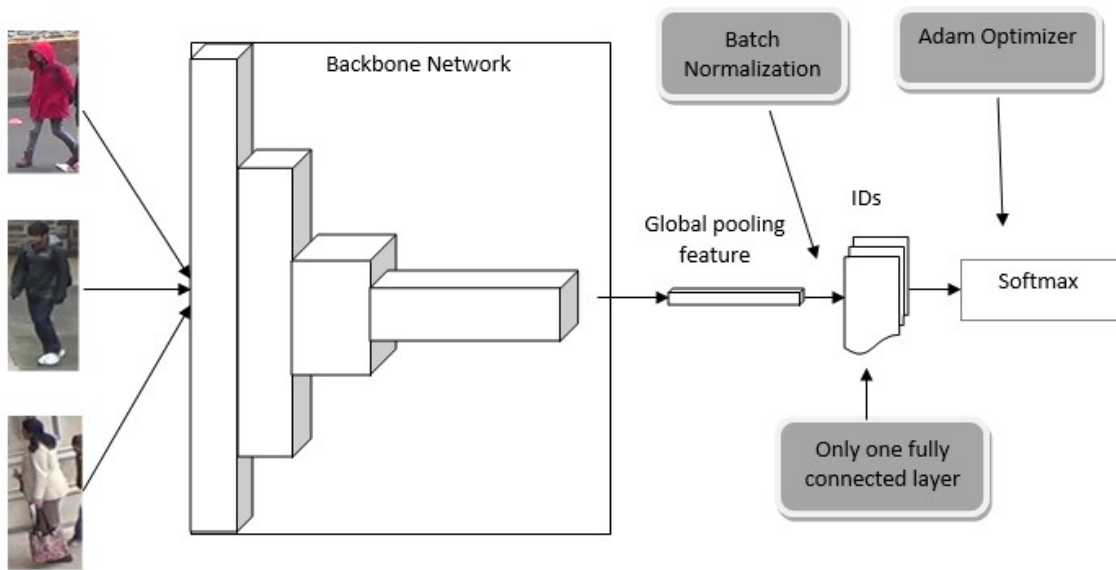


Fig.3.1 Pipeline of proposed CNN based approach

Based on experiments conducted, it can be said that these approaches are helpful in dealing with the problem of over-fitting and preserving the power of model which is trained on ImageNet. ImageNet dataset consists of more than 14 million images which provide many visual concepts. Many of the recent pre-trained models have shown the ability of good

transfer in different computer vision tasks. Therefore, we can design a good procedure of fine-tuning which can benefit the re-identification model by preventing the over-fitting and maintaining the power of pre-trained model. Following are the techniques for building an effective CNN baseline.

A. After global feature pooling use batch normalization

In order to effectively fine-tune a pre-trained (on ImageNet) CNN model for relatively small person re-identification datasets, one major issue is to reduce over-fitting during training. Currently, two widely used techniques for dealing with the problem of over-fitting are batch normalization [20] and dropout [21]. In dropout technique we randomly drop the outputs of some hidden neurons during the training. Batch-normalization is used to reduce the change in distribution of the network activations (internal covariant shift) which is caused by the change in parameters of the network during the training process. Output of each hidden neuron is normalized by using mean and variance of mini batch in batch normalization technique.

To prevent over-fitting, some works used dropout layer after global feature pooling. In proposed approach, instead of using dropout layer after global feature pooling we have used batch normalization. More stable gradient can be provided by batch normalization compared to dropout layer which randomly drops some hidden neurons. Generally, batch normalization shows the faster convergence and greater performance.

B. Use only one fully-connected layer

Usually there are 2 fully connected layers are used after global pooling layers in many common CNN models for person re identification. For dimensionality reduction of the feature space, first fully connected layer is used as bottleneck and for identity classification second fully connected layer is used. According to proposed approach, first fully connected layer which is used as bottleneck often reduces the final performance, bottleneck layer is not used in proposed method. This will help in easily back propagate the gradient from the softmax loss, this is also benefits in reducing the over-fitting.

C. Use Adam optimizer

Adam is a gradient based optimizer function which was recently proposed. Final practice in proposed approach in to use Adam optimization function to train CNN model. Parameters are updated according to the equation 3.1 in Adam:

$$w_t = w_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} \quad (3.1)$$

Where,

$$\hat{m}_t = \frac{m_t}{1 - \alpha_1^t} \quad (3.2)$$

$$\hat{v}_t = \frac{v_t}{1 - \alpha_2^t} \quad (3.3)$$

$$m_t = \alpha_1 m_{t-1} + (1 - \alpha_1) g_t \quad (3.4)$$

$$v_t = \alpha_2 v_{t-1} + (1 - \alpha_2) g_t^2 \quad (3.5)$$

Where, t is time step, η is learning rate, α_1 and α_2 and ϵ are hyper parameters.

Currently, SGD (stochastic gradient descent) is used in several state of the art methods. In comparison to stochastic gradient descent, Adam help in smoothing the variations in gradients. This will reduce the unessential disruption in model.

CHAPTER 4

EXPERIMENT

4.1 Experimental Setup

Datasets: Experiments were conducted on two widely used datasets for person re-identification: Market-1501 [22] and DukeMTMC-reID [23]. There are 32,668 images of 1501 person in the Market-1501 dataset and all these images are from six different cameras of different resolutions. In DukeMTMC-reID dataset there are images of 1,812 individuals with 8 cameras, from which 1,404 individuals appears in more than two cameras and 408 persons who appears in only one camera. For training only Market-1501 dataset is used while testing is conducted on both the datasets. For training set images of 751 persons are used and the images of remaining 750 persons are used for test set from maket-1501 dataset. Below figure shows the samples of images from these datasets:



Fig.4.1 Image samples from Market-1501 and DukeMTMC-1501 datasets, image pairs with green bounding box are matched pairs and with red bounding box are unmatched pairs.

Evaluation Metrics: For evaluating the person re-identification model, the metrics used are mean average precision (mAP) and Rank-n. mAP is calculated by finding the mean of average precision (AP) for a given query. AP is calculated as area under the Precision Recall (PR) curve. Precision and recall can be calculated by using eq. 4.1 and 4.2 respectively:

$$Precision = \frac{TP}{TP+FP} \quad (4.1)$$

$$Recall = \frac{TP}{TP+FN} \quad (4.2)$$

Where, TP (True Positive) is the total number of correct predictions for actual positive samples, FP (False Positive) is the total number of incorrect predictions for positive samples and FN (False Negative) is the number of incorrect predictions for negative samples.

Average Precision (AP) is calculated as the ratio of the sum of all the accuracies in a category to the number of the images in the category. Eq. 4.3 and 4.4 shows the steps for calculating the mean average precision:

$$AP = \frac{\sum precision}{C_i} \quad (4.3)$$

Where, C_i is the total number of the images in category i .

mAP is calculated as the mean of the all AP calculated for each category. Hence, mAP is used to evaluate all the categories.

$$mAP = \frac{\sum_{i=1}^N AP_i}{N} \quad (4.4)$$

Where N is the total number of categories.

Rank-n shows the probability that a query is present in different sized (n) lists. For example, Rank-5 represents the probability that query ID is present in first five results.

Implementation details:

- Software used: PyTorch Framework
- Backbone CNN model: Resnet50
- Resized input image: 256×128
- Augmentation: left-right image flipping
- Number of epochs: 60
- Optimizer: Adam (mini batch size = 32)
- Initial learning rate: 0.00035

Training: Market-1501 dataset used for the training. Learning rate plays an important role while training the person re-identification model. Initially learning rate is set to 3×10^{-4} .

For Market-1501 dataset fig 4.2 shows the loss and top-1 error vs epochs curve for training and validation.

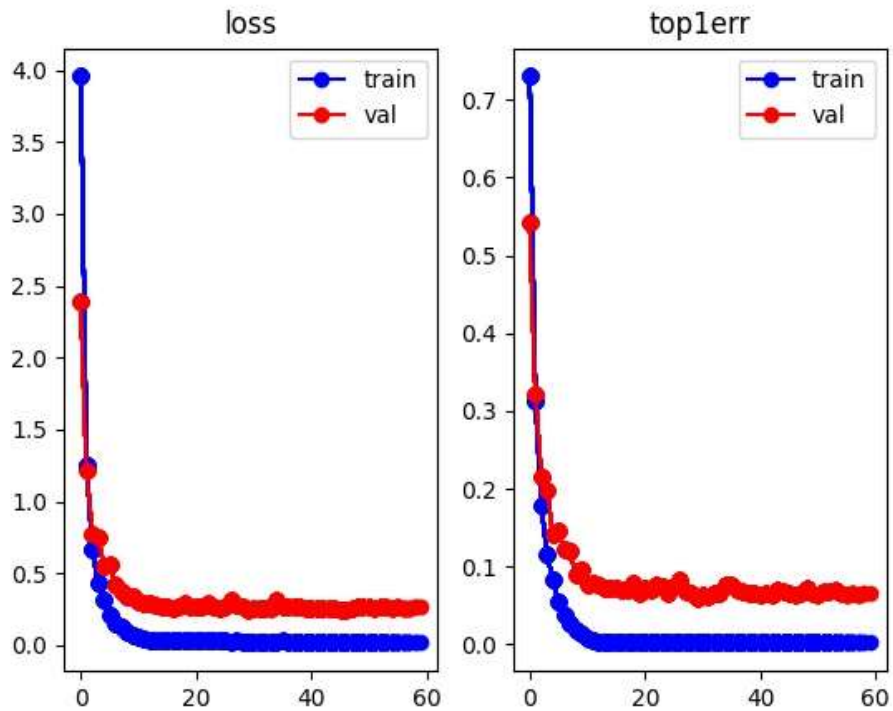


Fig.4.2 Training and validation loss curve and top1 error curve

Training and validation accuracy achieved in 60 epochs for market-1501 dataset is 98.5% and 96.2% respectively. Fig 4.3 shows the top-1 accuracy vs epochs curve for the training and validation.

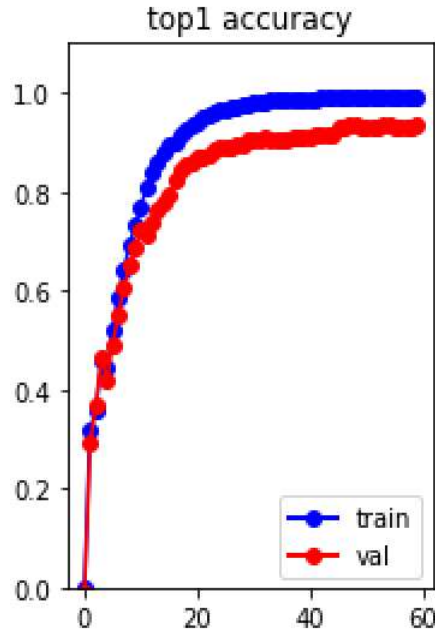


Fig 4.3 Training and validation top1 accuracy curve

4.2 Results

By using these three purposed approaches for fine tuning a pre-trained CNN model, experimental results have shown good performance for the Market-150 and DukeMTMC-reID datasets. The performance comparison of the purposed CNN baseline with some state-of-the-art methods which had used the pre-trained CNN models on ImageNet is shown in the following three tables for each one of three datasets.

Experimental results on Market-1501 dataset shows CMC score for rank-1 is 96.1 % and mean average precision of 93.3%, which is better than many state-of-the-art methods which have used fine-tuned CNN model pre-trained on ImageNet as a backbone network.

Table 3: Performance evaluation on Market-1501 dataset

Methods	mAP	Rank-1	Rank-5	Rank-10
CAP [24]	85.1	93.3	95.5	96.4
CBN+BoT [25]	83.6	94.3	95.9	96.7
ABD-Net [26]	88.28	95.6	96.0	96.3
Proposed Method	93.3	96.1	97.3	98.1

Performance comparison of proposed method with some state of the art methods on Market-1501 dataset

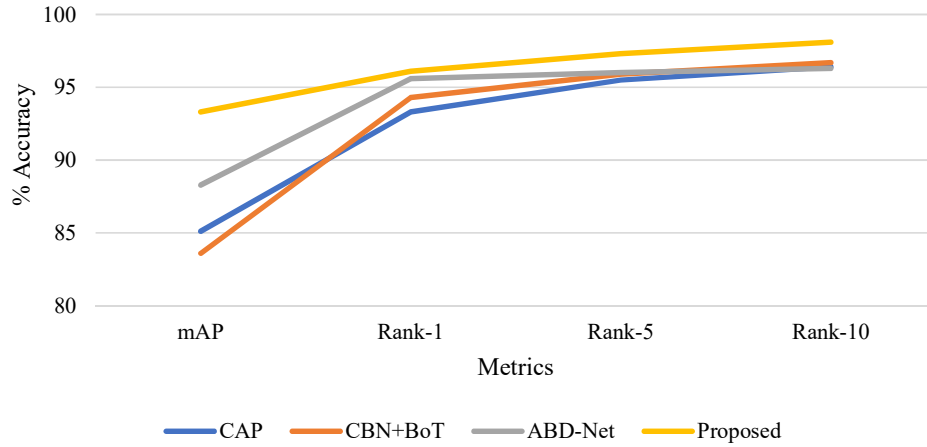


Fig. 4.3 Comparison with some state of the art methods.

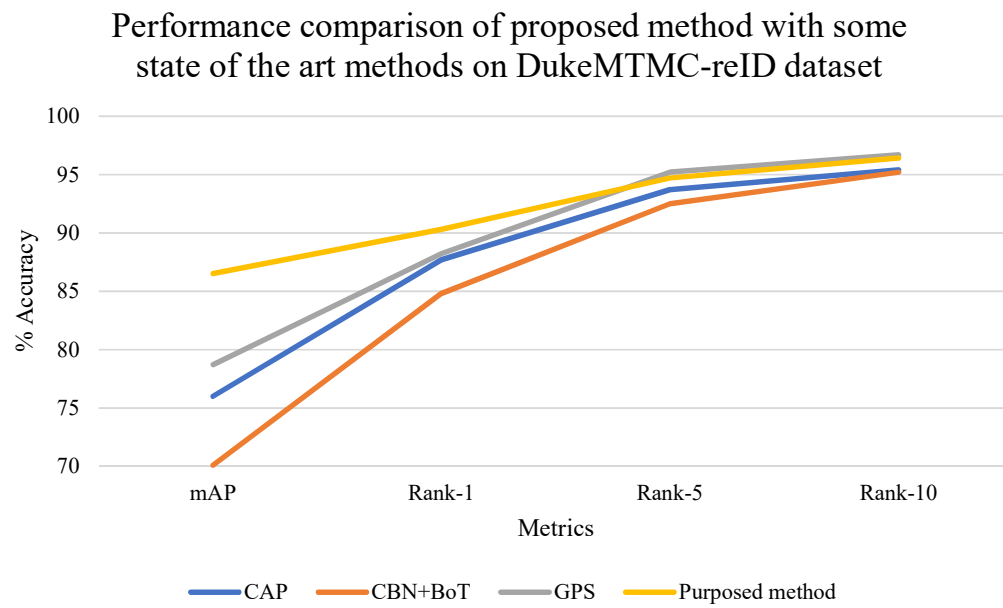
Results on DukeMTMC-reID dataset: Table 5 shows the comparison of performance of proposed method and some state of the art methods on DukeMTMC-reID dataset. From the table we can see that by using proposed approaches the performance of existing pre-trained CNN models can be significantly increased and can perform even better than state of the art methods. Performance on DukeMTMC-reID dataset is 90.3 and 86.5 for rank-1 and mAP respectively.

Table 5: Performance on DukeMTMC-reID dataset

Methods	mAP	Rank-1	Rank-5	Rank-10
CAP [24]	76	87.7	93.7	95.4
CBN+BoT [25]	70.1	84.8	92.5	95.2
GPS [27]	78.7	88.2	95.2	96.7
Proposed method	86.5	90.3	94.7	96.4

Comparing to the state of the art methods, proposed methods has shown better accuracy than some of the state of the art methods. Hence, we can conclude that without using any low level technical strategy or high level domain knowledge we can achieve the state of the art

performance by just following some rules for baseline building. Below is the plot for comparison.



CHAPTER 5

CONCLUSION

The objective of the thesis was to implement an effective CNN baseline using pre-trained network and set of rules that can improve the performance of pre-trained network and can deal with the problem of over fitting and can reduce it. In the experiment which was conducted on two benchmark datasets: Market-1501 and DukeMTMC-reID, we evaluated the performance of the pre-trained ResNet-50 which was fine-tuned using three practices which was described in chapter 3 and results shows that using these practices can effectively improve the performance and can deal with the problem of over fitting.

- Instead of dropout layer using batch normalization after global pooling layer helps in reducing over fitting.
- Removing the bottleneck layers improved the speed of training and made it easy to back propagate the gradient from the softmax loss.
- Adam optimizer is used as optimization function which is more suitable than SGD as it smoothen the variation in gradients.

Using above three techniques in our implementation, we have enhanced the performance of a simple pre-trained CNN model without making any high level changes and experimental results supports this argument.

5.1 Challenges

Even the purposed approach have shown good performance but still there are many failure cases due to some challenges like occlusion, similar outfit, low resolution of CCTV images. Also in some cases there may be more than one person in a single query image which can confuse the model. Some state-of-the-art methods have surpassed the human level accuracy but these challenges can even confuse a human brain, so this is near to impossible to achieve accuracy of 100% on all the datasets but it may be possible for some datasets.

5.2 Future work

In this work we have used only ResNet-50 to perform experiments and compared the results with methods which have used ResNet-50 as backbone network and surpassed the performance of many state-of-the-art methods on challenging datasets. As, when applying these techniques in ResNet-50 can result in enhanced performance, we should also apply these techniques on other CNN architectures which are pre-trained on ImageNet eg. GoogleNet, InceptionNet, DenseNet, ResNext, VGGNet etc. We suggest that the researchers should apply these proposed practices when they are using pre-trained CNNs for person re-identification in order to enhance the performance of backbone network.

REFERENCES

- [1] LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86, no. 11 (1998): 2278-2324.
- [2] Krizhevsky, Alex, Ilya Sutskever, and z Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012): 1097-1105.
- [3] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248-255. Ieee, 2009.
- [4] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).
- [5] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.
- [6] Bak, Sławomir, Etienne Corvee, Francois Bremond, and Monique Thonnat. "Person re-identification using spatial covariance regions of human body parts." In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 435-440. IEEE, 2010.
- [7] Shen, Yang, Weiyao Lin, Junchi Yan, Mingliang Xu, Jianxin Wu, and Jingdong Wang. "Person re-identification with correspondence structure learning." In *Proceedings of the IEEE international conference on computer vision*, pp. 3200-3208. 2015.
- [8] Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L., and Murino, V. (2011). Custom pictorial structures for re-identification. In *Bmvc*, volume 1, page 6. Citeseer.
- [9] Yang, Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li. "Salient color names for person re-identification." In *European conference on computer vision*, pp. 536-551. Springer, Cham, 2014.

- [10] Lin, Zhe, and Larry S. Davis. "Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance." In *International symposium on visual computing*, pp. 23-34. Springer, Berlin, Heidelberg, 2008.
- [11] Prosser, Bryan James, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q. Mary. "Person re-identification by support vector ranking." In *BMVC*, vol. 2, no. 5, p. 6. 2010.
- [12] Li, Wei, Rui Zhao, Tong Xiao, and Xiaogang Wang. "Deepreid: Deep filter pairing neural network for person re-identification." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 152-159. 2014.
- [13] Chang, Xiaobin, Timothy M. Hospedales, and Tao Xiang. "Multi-level factorisation net for person re-identification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2109-2118. 2018.
- [14] Sun, Yifan, Liang Zheng, Weijian Deng, and Shengjin Wang. "Svdnet for pedestrian retrieval." In *Proceedings of the IEEE international conference on computer vision*, pp. 3800-3808. 2017.
- [15] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9. 2015.
- [16] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700-4708. 2017.
- [17] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818-2826. 2016.
- [18] Zheng, Liang, Yi Yang, and Alexander G. Hauptmann. "Person re-identification: Past, present and future." *arXiv preprint arXiv:1610.02984* (2016).

- [19] Gou, Mengran, Ziyang Wu, Angels Rates-Borras, Octavia Camps, and Richard J. Radke. "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets." *IEEE transactions on pattern analysis and machine intelligence* 41, no. 3 (2018): 523-536.
- [20] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In *International conference on machine learning*, pp. 448-456. PMLR, 2015.
- [21] Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research* 15, no. 1 (2014): 1929-1958.
- [22] Zheng, Liang, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. "Scalable person re-identification: A benchmark." In *Proceedings of the IEEE international conference on computer vision*, pp. 1116-1124. 2015.
- [23] Zheng, Zhedong, Liang Zheng, and Yi Yang. "Unlabeled samples generated by gan improve the person re-identification baseline in vitro." In *Proceedings of the IEEE international conference on computer vision*, pp. 3754-3762. 2017.
- [24] Wang, Menglin, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. "Camera-aware Proxies for Unsupervised Person Re-Identification." *arXiv preprint arXiv:2012.10674* (2020).
- [25] Zhuang, Zijie, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. "Rethinking the distribution gap of person re-identification with camera-based batch normalization." In *European Conference on Computer Vision*, pp. 140-157. Springer, Cham, 2020.
- [26] Chen, Tianlong, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. "Abd-net: Attentive but diverse person re-identification." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8351-8361. 2019.
- [27] Nguyen, Binh X., Binh D. Nguyen, Tuong Do, Erman Tjiputra, Quang D. Tran, and Anh Nguyen. "Graph-based Person Signature for Person Re-Identifications."

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3492-3501. 2021.