國立臺灣大學電機資訊學院電信工程學研究所

碩士論文

Graduate Institute of Communication Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

(論文題目) 台大碩士論文範本

(英文題目) Example of NTU Master Thesis

林彥廷

Yan-Ting Lin

指導教授：傅立成 博士

Advisor: Li-Chen Fu, Ph.D.

中華民國110年1月

January, 2021

誌謝

回首過去的自己，碩班的這兩年來，有了太多的改變。從一開始對於影像處理沒有任何基礎認知到最後完成了一個研究。在這過程中很感謝大家的支持。

再來也感謝我的指導教授 傅立成教授，在這兩年來帶給我對於學術研究精神和態度的正確指引。老師致力於把學術研究的成果導向日常生活應用的精神也讓我學習到了很多實作的技巧以及當遇到問題時解決的應變能力，謝謝老師。

此外也要感謝在百忙中撥空來參與我的口試考試的口試委員：傅楸善教授、張文中教授、黃正民教授以及王鈺強教授，在口試過程中對於研究給的各種意見以及如何把碩士論文詮釋得更完整給了相當多的意見，這些意見也讓也讓我也讓我的碩士論文更加完整。也感謝影像組碩二的夥伴們，在研究的過程中我們遇到了許多挫折以及困境，大家一起互相給意見、彼此分享技術；也謝謝電機ACL其他夥伴，這兩年大家共同經歷了許多的展示以及會議，大家一起學習、成長還有玩樂。

另外也謝謝學長在這兩年給了我許多的協助還有指導，幫影像組減輕了不少的負擔，讓我們可以減輕許多的壓力把更多的精力專注在研究上，在我許多需要幫忙的時候都二話不說馬上答應我，在未來的研究路上還會有很多的挑戰，大家可以在這過程中得到許多經驗也相信各位可以在未來運用這些經驗來解決研究上的問題。

Tom January 11, 2021

1. 中文摘要

近年來，人物重新識別系統受到大量的關注，因為其有廣大的應用場域，像是智慧家庭、健康照護以及監視系統。但是隨著視角的改變以及拍攝相機的位置不同，人的輪廓外觀也會跟著不同，這造成了從不同的視角進行行人追蹤仍然是個挑戰。基於上述提到的挑戰，這篇論文設計了一個能夠在應用在實際場域的人物重新識別深度學習模型，具有較高的準確度同時兼顧了模型的複雜度。

對於背景所造成的雜亂，這篇論文提出了一個新的加權池化方法，能夠聚合特徵圖上所有響應高的特徵並且抑制背景區域響應低的特徵，加權池化能夠改善平均池化與最大池化的缺點，有效的剔除對於判別人物無益的背景雜亂，也考量到了人物身上會具有多個具有判別性的特徵並加以整合。

**關鍵字：**深度學習、資料檢索、人物重新識別、聚合損失函數

1. ABSTRACT

In recent years person re-identification (Re-ID) [1] has raised lots of attention in the area of computer vision. It has a wide range of applications including smart home, elderly care and surveillance systems. In general, Re-ID is challenged by background clutter, occlusion, different camera viewpoints and identities with a similar human appearance. The shape of the human body looks completely different from different viewpoints. hence tracking humans from different cameras remains a challenging problem. These factors hinder the process of extracting robust and discriminate representations.

For tackling the background clutter issue, we propose a new weighted pooling (WP) that takes advantage of average pooling and max pooling. WP can enlarge the difference of response value between salient points and unimportant regions and then aggregates the pixels with high response. Different from the global average pooling, the unwanted background information will not be fed into the followed fully-connected layer. At the same time, weighted pooling can spatially merge all the important points in a feature map for final prediction, which is different from max pooling.

Recent studies have proved the effectiveness of attention mechanisms among different tasks including machine translation, object detection and image classification. In this thesis, we design a non-local attention method that can guide the model to focus on interesting regions by capturing long-range dependencies.

First of all, to solve the problem of brightness changes, the most intuitive way is to collect an even larger person re-identification dataset subject to various brightness levels, which however is very expensive to collect and label. Therefore, this thesis proposes an illumination-invariant feature vector that assists in model training based on synthetic data. To remove the shortcomings of the time complexity of the metric loss function, we propose the clustering-based loss function to reduce the time complexity, and we also show that the performance of the proposed loss function is better than the metric loss function. In the final experiment, it is also proved that the proposed method in this thesis excels the state-of-the-art methods of resolving person re-identification problems.

**Keywords:** Deep learning, Information retrieval, Person re-identification

1. CONTENTS

口試委員會審定書 #

[誌謝 i](#_Toc59930317)

[中文摘要 ii](#_Toc59930318)

[ABSTRACT iii](#_Toc59930319)

[CONTENTS v](#_Toc59930320)

[LIST OF FIGURES vii](#_Toc59930321)

[LIST OF TABLES ix](#_Toc59930322)

[Chapter 1 Introduction 1](#_Toc59930323)

[1.1 Motivation 1](#_Toc59930324)

[1.2 Literature Review 3](#_Toc59930325)

[1.2.1 Human Detection 3](#_Toc59930326)

[1.2.2 Person Re-Identification 5](#_Toc59930327)

[1.3 Contributions 6](#_Toc59930328)

[1.4 Thesis Organization 6](#_Toc59930329)

[Chapter 2 Preliminaries 7](#_Toc59930330)

[2.1 Convolutional Neural Networks (CNNs) 7](#_Toc59930331)

[2.2 Residual Network 11](#_Toc59930332)

[2.3 Objective Functions 12](#_Toc59930333)

[2.3.1 Cross Entropy Loss 12](#_Toc59930334)

[2.3.2 Triplet Loss with PK Batch and Batch Hard 13](#_Toc59930335)

[2.3.3 Center Loss 15](#_Toc59930336)

[2.4 Information Retrieval 15](#_Toc59930337)

[2.5 Re-ranking Algorithm 17](#_Toc59930338)

[2.6 Knowledge Distillation 18](#_Toc59930339)

[Chapter 3 Person Re-Identification 19](#_Toc59930340)

[3.1 System Overview 19](#_Toc59930341)

[3.2 Knowledge Distillation Re-ID 20](#_Toc59930342)

[3.2.1 Knowledge Formulation 20](#_Toc59930343)

[3.2.2 Knowledge Receiver 22](#_Toc59930344)

[3.3 Non-local Attention 23](#_Toc59930345)

[3.4 Re-weighted Average Pooling (RAP) 23](#_Toc59930346)

[3.4.1 Pros and Cons of GAP and GMP 23](#_Toc59930347)

[3.4.2 The Effectiveness of RAP 23](#_Toc59930348)

[3.5 Training and Inference Details 25](#_Toc59930349)

[Chapter 4 Experiments 26](#_Toc59930350)

[4.1 Configuration 26](#_Toc59930351)

[4.2 Training Details 27](#_Toc59930352)

[4.3 Person Re-identification Datasets 27](#_Toc59930353)

[4.3.1 Market-1501 Dataset 28](#_Toc59930354)

[4.3.2 DukeMTMC-reID Dataset 29](#_Toc59930355)

[4.3.3 Evaluation Metrics 30](#_Toc59930356)

[4.4 Experimental Results 33](#_Toc59930357)

[4.5 Ablation Studies 33](#_Toc59930358)

[REFERENCE 34](#_Toc59930359)

1. LIST OF FIGURES

[Figure 1‑1. Each row represents the images of the same identity from different camera viewpoints. 2](#_Toc59930293)

[Figure 1‑2. The background subtraction result [10] 3](#_Toc59930294)

[Figure 1‑3. The architecture of FairMot [12] 4](#_Toc59930295)

[Figure 1‑4. The purpose of person Re-ID 5](#_Toc59930296)

[Figure 1‑5. Spatial and channel partition representation network [14] 5](#_Toc59930297)

[Figure 2‑1. A standard image classification CNN architecture contains several convolutional layers and FC layers. 7](#_Toc59930298)

[Figure 2‑2. A standard convolution operation with stride equals 1. 8](#_Toc59930299)

[Figure 2‑3. The configurations of AlexNet, VGG16 and Inception Net. 10](#_Toc59930300)

[Figure 2‑4. Left: a building block for ResNet34. Right: a “bottleneck” building block for ResNet-50. 11](#_Toc59930301)

[Figure 2‑5. The family of ResNet [13]. The model is named by the number of layers it has. 12](#_Toc59930302)

[Figure 2‑6. Anchor, positive sample and negative sample. 13](#_Toc59930303)

[Figure 2‑7. The distribution of deeply learned features under the supervision of (a) ID loss and (b) ID Loss with center loss. 15](#_Toc59930304)

[Figure 2‑8. The algorithm of re-ranking for person Re-ID[28]. 17](#_Toc59930305)

[Figure 3‑1. The framework of the proposed Re-ID model. 19](#_Toc59930306)

[Figure 3‑2 22](#_Toc59930307)

[Figure 4‑1. The different identities collected from different cameras [34]. 28](#_Toc59930308)

[Figure 4‑2. Sample images of the distractor dataset (junk) [34]. 28](#_Toc59930309)

[Figure 4‑3. The camera position setting for DukeMTMC-re-ID. 29](#_Toc59930310)

[Figure 4‑4. The images distribution on DukeMTMC-reID training set. 30](#_Toc59930311)

[Figure 4‑5. The Cumulative Match Characteristic (CMC) curve. 31](#_Toc59930312)

[Figure 4‑6. Rank- calculation. 31](#_Toc59930313)

1. LIST OF TABLES

[Table 2‑1. Confusion Matrix 16](#_Toc59930260)

[Table 4‑1. Specification of Environment 26](#_Toc59930261)

# Introduction

In this chapter, we first describe the motivation of this work in Section 1.1. A complete literature review is presented in Section1.2. In Section1.3, we highlight our contributions, and the organization of this thesis is presented in Section1.4.

## Motivation

Recently, person re-identification(Re-ID) [1] has raised lots of attention in the area of computer vision, due to its wide applications, such as elderly care, smart home, and surveillance systems. The objective of person re-identification is to associate individual identities across different cameras, timing can locations. Given a person-of-interest (query) image, the person Re-ID model will embed each image into feature space and then rank the similarities of the feature embeddings between the gallery images and the query. The goal is to retrieve the images in gallery with identity same to query. Person Re-ID is inherently challenging because of the significant visual appearance changes caused by various factors such as background clutter, occlusion, different camera viewpoints, identities with similar human appearance and posture variation.

Over the last decade, there are some well-known approaches that utilize hand-crafted features, such as Histogram of Oriented Gradient (HOG) [2] and Scale-Invariant Feature Transform (SIFT) [3] feature. However, these kinds of traditional methods suffer from cross-camera person re-identification problem and their performances drop sharply when the viewpoint changes. This is because hand-craft features are view dependent but the appearance of a person is continuous changing, as shown in Figure 1‑1. Thus, the misalignment problem becomes one of the most critical issue.

It has become a general fact that, deep learning methods can obtain more effective

|  |
| --- |
|  |
| Figure 1‑1. Each row represents the images of the same identity from different camera viewpoints. |

features in different tasks since the learned models can extract high-level representative feature from data through training instead of the hand-crafted features. Since 2010, Felzenszwalb *et al*. [4] proposed a novel method combining the HOG [2] feature and Support Vector Machine (SVM) [5], and thus machine learning methodology go beyond the traditional hand-crafted feature method in VOC competition [6]. Due to the development of powerful learning-based human detector, most researches design the Re-ID network which only focuses on appearance feature extraction.

In our prior knowledge, face recognition has already been developed. However, person re-identification technology is still hard to apply to real world system due to the high computational cost which is caused by the deep learning model. And the feature extracted by the model cannot handle the aforementioned challenges from the complex real-world environment. Another key issue is the non-negligible domain gap between real-world scenario and training dataset. In this thesis, we aim to build a deep learning model that can withstand the aforementioned challenges to fulfill the goal of Re-ID task.

## Literature Review

We first give a brief review of human detection method in Section 1.2.1, then discuss the algorithm of person re-identification in Section 1.2.2.

### Human Detection

Due to the need of various applications in computer vision, human detection has raised lots of attention lately. According to the input data, human detection can be categorized into two aspects: (1) depth image-based human detection and (2) RGB image-based human detection.

For depth image, some methods [7-9] reveal the foreground images by mean of background subtraction. And a classifier is added to achieve human detection. Figure 1‑2 shows the background subtraction result. Background image initialization is required before performing background subtraction. The foreground image (see Figure 1‑2(c)) is generated by element-wise subtraction of the whole image (see Figure 1‑2(a)) from the background image (see Figure 1‑2(b)).

|  |  |  |
| --- | --- | --- |
|  |  |  |
| (a) whole image | (b) background image | (c) foreground result |
| Figure 1‑2. The background subtraction result [10] | | |

The above method is also suitable for handling RGB images. However, background subtraction cannot be applied to cluttered and mutative environment, which elevates the difficulty when it comes to application. Dalal *et al.* [11] proposed HOG-based human detection method which can detect human even in cluttered backgrounds under varying illumination. Felzenszwalb *et al*. [4] trained multiple models to solve the problem where features are difficult to be extracted due to perspective change and has a significant improvement over [11].

|  |
| --- |
|  |
| Figure 1‑3. The architecture of FairMot [12] |

There has been remarkable progress on object detection and re-identification (re-ID) in recent years which are the key components of multi-object tracking. FairMOT [12] is an effective bottom-up multiple object tracking algorithm which consists of an encoder-decoder network and two homogeneous branches for detecting objects and extracting re-ID features, respectively. The authors adopt ResNet-34 [13] as backbone of encoder-decoder in order to strike a good balance between accuracy and speed. The detection branch is implemented in an anchor-free style which estimates object centers and sizes represented as position-aware measurement maps. In this thesis, we utilize the human detection result of FairMOT combined with the proposed person Re-ID framework that has powerful representative ability to construct an online person Re-ID system.

### Person Re-Identification

|  |
| --- |
|  |
| Figure 1‑4. The purpose of person Re-ID |

With the absurd development of deep neural network (NN) and increasing demand of intelligent surveillance system, it has gained significantly increased interest in the computer vision community and has achieved inspiring success on a number of datasets. Given a person query image, person Re-ID aims at associating images among multiple non-overlapping cameras with the same identity in the gallery, as shown in Figure 1‑4.

|  |
| --- |
|  |
| Figure 1‑5. Spatial and channel partition representation network [14] |

## Contributions

## Thesis Organization

# Preliminaries

Some prerequisite knowledge about this work will be introduced in this chapter. We will give the background of deep NN, including convolutional layers and residual network in section 2.1 and 2.2, followed by some typical objective functions like cross entropy loss, triplet loss and center loss in section 2.3.

## Convolutional Neural Networks (CNNs)

Since AlexNet [15] first applied CNN to accomplish the task of image classification and achieved impressive performance in ImageNet [16], CNN has made a considerable amount of success in computer vision, including, image classification [15, 17], objection detection [10, 18] and action recognition[19, 20], etc.

|  |
| --- |
|  |
| Figure 2‑1. A standard image classification CNN architecture contains several convolutional layers and FC layers. |

In the following paragraph, we use image classification as an example to simplify the explanation. Traditional frameworks consist of hand-crafted feature extractor like SIFT [3] or HOG [2] and learnable classifiers such as Support Vector Machine (SVM) [5]. Though solving the objective function can adjust the parameters of classifiers, pre-defining the initial parameters of feature extractor is needed, which requires some domain knowledge and several trial and error for searching the most suitable hyper-parameters of the classifier.

In contrast, the CNN-based frameworks consist of learnable feature extractors and classifiers. That is, both the feature extractor and classifier can learn by back-propagation in an end-to-end manner, making it easier to extract reliable features. Figure 2‑1 shows a standard CNN architecture for the classification task.

|  |
| --- |
|  |
| Figure 2‑2. A standard convolution operation with stride equals 1. |

Traditional NNs operate in a node-to-node manner. By contrast, convolution is using a ‘kernel’ to extract certain ‘features’ from an input image which takes spatial information into consideration. A kernel is a matrix, which is slid across the image and multiplied with the input such that the output is enhanced in a certain desirable manner. Each convolutional layer has multiple kernels that perform dot products with a specific input patch, generating different output in channel (or depth) dimension, see Figure 2‑2. Zeiler *et al*. [17] introduced Deconvolution to visualize the feature map response and showed that shallow layers extract low-level features such as edge, corner and color. The layers in the deep are responsible for retrieving high-level features.

In 2014, Simonyan *et al*. proposed VGG[21], which is deeper than AlexNet [15] and achieves higher accuracy. Inception Net [22] was proposed by Google Inc. in the same year. The concept of Inception is widening the structure which is not like VGG and AlexNet with deeper structure. It requires fewer parameters and reaches higher performance. Figure 2‑3 shows the configurations of AlexNet, VGG16 and Inception Net.

|  |
| --- |
|  |
| (a)AlexNet |
|  |
| (b)VGG16 |
|  |
| (c)Inception blocks |
| Figure 2‑3. The configurations of AlexNet, VGG16 and Inception Net. |

## Residual Network

|  |
| --- |
|  |
| Figure 2‑4. Left: a building block for ResNet34. Right: a “bottleneck” building block for ResNet-50. |

According to Universal Approximation Theorem [23], the feed-forward neural network is capable to closely fit any non-linear equation mapping. Researchers start to deepening the model for strengthening the generalization ability. However, the vanishing gradient problem [24] occurs when training a deep model with gradient-based learning methods and backpropagation. In such methods, each of the neural network's weights receives an update proportional to the partial derivative of the objective function with respect to the current. The problem is that in some cases, the gradient will be vanishingly small, effectively preventing the weight from changing its value. One possible method is replacing the activation function with saturation region, such as Hyperbolic Tangent and Sigmoid, by Rectified Linear Unit (Relu).

Another effective solution is proposed by Kaiming *et al* [13]. They introduce the Residual Network (ResNet) which is a feedforward neural network by utilizing the framework called shortcut (skip) connection. This novel design makes a very deep network trainable. Figure 2‑4 illustrates the configuration of residual building block and bottleneck.

The ResNet50, shown in Figure 2‑5, has 50 layers formed into five stages. With the powerful generalization ability, it is often employed as the backbone of deep learning models. In this thesis, we adopt a revised ResNet50 as the backbone and further details will be revealed in Chapter 3.

|  |
| --- |
|  |
| Figure 2‑5. The family of ResNet [13]. The model is named by the number of layers it has. |

## Objective Functions

In this section we are going to discuss three loss functions used in this thesis.

### Cross Entropy Loss

Also called Softmax Loss or ID Loss, Cross Entropy Loss is commonly used in machine learning. Cross-entropy is a measure from the field of information theory, building upon entropy and generally calculating the difference between two probability distributions. Cross-entropy loss measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverge from the actual label. Predicting a probability of .012 when the actual observation label is 1 would be bad and result in a high loss value. A perfect model would have a log loss of 0. In binary classification, where the number of classes equals 2, cross-entropy can be calculated as:

|  |  |
| --- | --- |
|  | 2‑1 |

For the multiclass classification task, the loss is:

|  |  |
| --- | --- |
|  | 2‑2 |

where *C*, and denotes the number of categories, ground truth and the predicted probability. Given a sample, the model will produce a corresponding logit for each category. Then a softmax function will convert those logits into probabilities:

|  |  |
| --- | --- |
|  | 2‑3 |

### Triplet Loss with PK Batch and Batch Hard

Deep Metric Learning (DML) aims at learning a distance function that measures the similarity between samples in the feature space. Triplet Loss [25] a commonly used loss function to help converge the model. The basic idea is that contextually similar data points are projected in the near-by region whereas dissimilar data points are projected far away from each other in the high dimensional feature space.

|  |
| --- |
|  |
| Figure 2‑6. Anchor, positive sample and negative sample. |

For Triplet Loss, the objective is to build triplets <Anchor, Positive, Negative> consisting of an anchor image, a positive image (which is similar to the anchor image) and a negative image (which is dissimilar to the anchor image). We denote features of Anchor, Positive and Negative as , and .

Triplet Loss aims as minimizing <Maximizing> Euclidean distance between and <>.

|  |  |
| --- | --- |
|  | 2‑4 |
|  | 2‑5 |

**The objective is to learn representations with a small distance between them for positive pairs and greater distance than some margin value   for negative pairs.** When utilizing triplet loss to enhance feature discriminability, the way to sample images to form a triplet is critical. Hermans *et al*. [26] propose batch for sampling and batch hard for training. The core idea of batch is to form batches by randomly selecting identities and each with samples, resulting in a batch of images. Batch hard aims at choosing the hardest positive (negative) pair with the largest (smallest) distance. Base on the two tricks, we can rewrite Equation 2‑4:

|  |  |
| --- | --- |
|  | 2‑6 |

where data point corresponds to the *a*-th image of -th person in the mini-batch. This results in terms contributing to the loss. Additionally, the selected triplets can be considered moderate triplets since they are the hardest within a small subset. With triplet loss, we can regularize the feature produced by the deep learning model to be discriminative.

### Center Loss

The data distribution in the feature space will be like Figure 2‑7(a) if the model is trained under the supervision of ID loss only. Though the sample points are separable, there are considerable intra-class variations between them. Yandong et al. [27] develop Center Loss to further reduce the intra-class variations effectively. The purpose of center loss is to minimize the intra-class variations and it is defined as:

|  |  |
| --- | --- |
|  | 2‑7 |

The is the -th class canter of deep features. It is worth noting that during back-propagation, only the samples within the sample class in the mini-batch will be used to update the corresponding class center. In Figure 2‑7(b), we can see that the distribution of feature points is more separated than that in Figure 2‑7(a).

|  |  |
| --- | --- |
| 1. ID loss | 1. ID loss + center loss |
| Figure 2‑7. The distribution of deeply learned features under the supervision of (a) ID loss and (b) ID Loss with center loss. | |

## Information Retrieval

In this section we will give a brief introduction to Information Retrieval (IR) including definition, evaluation metric and how it can be used in the Re-ID problem.

IR is a typical task and has a variety of applications such as search engines, media search and digital libraries. Information retrieval (IR) is the activity of obtaining information system resources that are relevant to an information need from a collection of those resources. An information retrieval process begins when a user enters a query into the system. Re-ID is a subtask of IR. Given a person-of-interest (query) image, the person Re-ID model will embed each image into feature space and then rank the similarities of the feature embedding between the gallery images and the query. The goal is to retrieve the images in the gallery with identity same to query.

|  |
| --- |
| Table 2‑1. Confusion Matrix |
| |  |  |  |  | | --- | --- | --- | --- | |  | | Prediction | | | True | False | | Ground Truth | True | True Positive (TP) | False Negative (FN) | | False | False Positive (FP) | True Negative (TN) | |

The most commonly used evaluation metric in IR is Mean Average Precision (mAP). Before that, we will do a quick on the meaning of precision and recall. Precision is the fraction of relevant instances among the retrieved instances, while recall is the fraction of retrieved relevant instances among all relevant instances. In other words, Precision measures how accurate your predictions are and Recall measures how well you find all the positives. Here is the mathematical definition:

|  |  |
| --- | --- |
|  | 2‑8 |
|  | 2‑9 |

where , and denote true positive, false positive and false negative (see Table 2‑1). The method of calculating average precision and mean average precision will be elaborated in 4.3.3.

## Re-ranking Algorithm

When considering person Re-ID as a retrieval process, re-ranking is a critical step to improve its accuracy. Yet in the re-ID community, limited effort has been devoted to re-ranking, especially those fully automatic, unsupervised solutions. Thus Zhong *et al*. [28] propose a -reciprocal encoding method to re-rank the Re-ID results(see )

|  |
| --- |
|  |
| Figure 2‑8. The algorithm of re-ranking for person Re-ID[28]. |

The authors assume that if a gallery image is similar to the probe query in -reciprocal nearest neighbors then it is more likely to be the right match. More specifically, we take one gallery image as the probe query to the others. Once the probe query is ranked top- in similarity list the more likely the gallery image a true match is. For a fixed image, the -reciprocal nearest neighbors are encoded as a single vector to calculate the the -reciprocal feature. The idea is that top- images should belong to the same and the feature of different images will provide abundant information. The similarity between vectors is measured by Jaccard Distance, a distance to measure the difference between the two sets index.

The origin and the Jaccard distance are combined to describe the similarity between two sets. The re-ranking method does not require any labor work, so it is suitable for large-scale data sets.

## Knowledge Distillation

In machine learning, knowledge distillation (KD) [29] is the process of transferring knowledge from a large model to a smaller one. The small model is trained to mimic larger model. This training setting can be referred to as “teacher-student”, where the large model is the teacher and the small model is the student.

If both models are trained by the same supervision (label), the small model may have insufficient capacity to learn the information as rich as the large one. We believe the predicted score from the larger model contain useful knowledge representation somehow. For instance: when a model correctly predicts a class, it assigns a large value to the output variable corresponding to such class, and smaller values to the other output variables. The distribution of values among the outputs provides information on how the large model represents knowledge.

# Person Re-Identification

In this chapter, the overall architecture and algorithm will be elaborated. We first give an overview of the proposed framework in section 3.1, followed by knowledge distillation Re-ID in section 3.2, including knowledge formulation and knowledge receiver (KR). And the operation and meaning of non-local attention (NLA) will be mentioned. Then the proposed re-weight average pooling (RAP) will be introduced in section 3.4, including the mathematical prove of its effectiveness. We also explain the pros and cons of global average pooling (GAP) and global max pooling (GMP). Finally, training and inference details will be discussed.

## System Overview

|  |
| --- |
|  |
| Figure 3‑1. The framework of the proposed Re-ID model. |

The proposed Re-ID model is illustrated in Figure 3‑1. We believe the deeper layers can produce features that are more discriminative and informative than that of shallow layers. Hence, we propose a self-distilled framework that can distill knowledge within network itself. The networks are firstly divided into several sections. Then the knowledge in the deeper portion of the networks is squeezed into the shallow ones. We also integrate the modified non-local attention block into the network in order to aggregate semantically similar pixels in the spatial domain. With the aid of non-local attention, long-range (global) dependencies of three-dimensional tensors can be captured. Different from others utilizing GAP or GMP to reduce the size of features spatially in the second last layer, we design a new pooling method called “RAP” that can get rid of the background clutters and preserve several salient (discriminative) pixels for final prediction.

## Knowledge Distillation Re-ID

In this thesis, we do not train the network by conventional KD training scheme. In that way, both teacher and student model should be loaded into the GPU-memory, which is computation and capacity costly. Instead, the network is trained by self-distillation manner.

We believe the deeper layers can produce features that are more discriminative and informative than that of shallow layers. Hence, a self-distilled framework is proposed.

### Knowledge Formulation

Given a sample, the deepest layer in the network will produce corresponding logit for each category ( for Market1501 dataset), as mentioned in Section 2.3.1. Then a softmax function will convert the logits into probabilities:

|  |  |
| --- | --- |
|  | 3‑1 |

where *C* and refers to the number of categories and the predicted probability. Then the typical Cross Entropy Loss (ID Loss) is computed as:

|  |  |
| --- | --- |
|  | 3‑2 |

where refers to the one-hot encoded ground true. When training the network, the parameters will be updated to make the drop. That being said, the network will learn how to produce a distribution to be as close as to the , which is one-hot.

As mentioned before, the scores produced by the deepest layer in the network are informative and rich in knowledge. Here we take image classification problem for a instance. Given an image of tiger with ground truth , it is quite possible the network will produce for each category (see Figure 3‑2 (b)). The ground truth tells the model that the animal in this image is a tiger and the model says that the animal is most likely to be a tiger (), which is correct. But apart from that, the model also tells us tigers are similar to lions, a least more than snake. We can also know that sharks are similar to whales and Snakes are not alike to the other four species, by observation on the predicted probabilities.

|  |  |
| --- | --- |
|  |  |
|  |  |
| Figure 3‑2. The One-hot label and soft label with different temperatures. | |

Knowing that the model can produce features that are not only discriminative but also rich in knowledge, we are going to distill the knowledge out as soft labels to supervise the shallow layers. The definition of soft label is:

|  |  |
| --- | --- |
|  | 3‑3 |

where is a temperature and when we get the standard softmax function. As grows, the probability distribution generated by softmax function becomes softer providing more information (see Figure 3‑2 (c) and (d)). In this way, the knowledge can be distilled out by a high temperature and is represented by soft label. The shallow layers are trained under the supervision of both soft label and hard label, which leads to an obvious higher accuracy supported by experiments results.

### Knowledge Receiver

## Non-local Attention

## Re-weighted Average Pooling (RAP)

### Pros and Cons of GAP and GMP

### The Effectiveness of RAP

|  |
| --- |
|  |
| Figure1 An easy example of CNN operation |

Figure1 shows how a convolutional filter works on a feature map. There are 2048 convolutional filters operating on the feature map, producing a feature cubic with 2048 channels. The proposed weighted pooling (WP) will consume the feature cubic, producing a 1-dimensional tensor with length of 2048 for final prediction. The equation of WP is shown below:

Here, we want to know the change in over change in . We can easily calculate:

|  |
| --- |
|  |
| Figure2 A fully-connected layer |

In general, we will apply a fully-connected layer on the 2048 dimensional embedded to make the final prediction. An easy example is shown In Figure2. The fc layer produces three scores, , and .

Once getting the three scores, a softmax function is used to convert them into probability value.

We denote as the probability that the sample is classified as category1.

The typical cross entropy loss is defined as:

denotes number of class, is the predicted probability of the -th element of -th image and is the ground-truth label of -th image. If the sample belongs to category1, the corresponding is written as:

Once getting the magnitude of , backpropagation will compute the [gradient](https://en.wikipedia.org/wiki/Gradient) with respect to each learnable weight in the model and the weights will be updated toward direction to decrease the loss.

where is the learning rate.

(we focus in this example).

## Training and Inference Details

# Experiments

In this chapter, we will first introduce the environment setting in section 4.1, followed by the implementation details in section 4.2 and two famous person Re-ID datasets in section 4.3. Then, the comparison of experimental results with state-of-the-art will be elaborated in section 4.4. Finally, ablation studies in section 4.5 will demonstrate the effectiveness of the proposed method.

Evaluating the performance of a Re-ID model can be conducted by means of image retrieval. We will rank all the images in the gallery by their distance (either cosine or L2) to the query in feature space in an ascending way. Those image with the same identity of query should be at the top of the list.

## Configuration

Table 4‑1 list the specification of our experiment setup. In this thesis, we utilize Pytorch as the Application Programming Interface (API) to build up deep learning model. The proposed model is trained on a personal computer equipped with NVIDIA GeForce GTX 1080ti GPU with 11G memory.

|  |
| --- |
| Table 4‑1. Specification of Environment |
| |  |  | | --- | --- | | Central Processing Unit (CPU) | AMD 3600x | | Graphic Processing Unit (GPU) | NVIDIA GeForce GTX 1080ti | | Random Access Memory (RAM) | 24.0 GB | | Operating System (OS) | Ubuntu 18.04 | | Deep Learning API | Pytorch 1.4.0 | |

## Training Details

We adopt ResNet-IBN [30] with pre-trained weights on ImageNet [16]. The size of input images is set to be . The RGB channels of input images are normalized by mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225], respectively. For the batch construction, we randomly choose 16 identities and each with 4 images to form a minibatch. The data augmentation technique such as random crops, horizontal flips, padding and random erasing [31] to consolidate the Re-ID model.

The learning rate is adjusted by warm up [32] strategy to bootstrap the network for better performance. Totally there are 115 training epochs. The learning rate is increased linearly from to . Then, the learning rate is decayed to and at 40-th epoch and 70-th epoch respectively. The learning rate at epoch is computed as:

|  |  |
| --- | --- |
|  |  |

The training set and testing set of person Re-ID do not overlap, the label smooth technique [33] is utilized to prevent overfitting.

To enrich the granularity of feature, the last spatial downsampling operation in the backbone network is removed to increase the size of the feature map. The balance parameter and the margin of triplet loss are set to be 0.005 and 0.3, respectively.

## Person Re-identification Datasets

We evaluation the proposed architecture on two public dataset Market1501 [34] and DukeMTMC-reID [35] dataset, both of which are the most famous and challenging and will. The two datasets will be introduced in section 4.3.1 and 4.3.2. They contain a large number of images of different identities captured from several non-overlapping cameras. Evaluation metrics such as Cumulative Match Characteristic (CMC) curve and mAP will be mentioned in the end.

### Market-1501 Dataset

The Market-1501 dataset [34] was collected on the campus of Tsinghua University, China. There are total 1501 identities (pedestrians) captured by six cameras and 32668 detected bounding boxes. Each identity is captured by at least two cameras.

|  |
| --- |
|  |
| Figure 4‑1. The different identities collected from different cameras [34]. |
|  |
| Figure 4‑2. Sample images of the distractor dataset (junk) [34]. |

There are 751 identities with 12936 images in the training set and 750 identities with 19732 images in the testing set. The pedestrian bounding box of 3368 query images was drawn manually while the pedestrian bounding in the gallery was detected by Deformable Part Model (DPM) [4]. Figure 4‑1 demonstrates some images in Market-1501. As we can see in the second row in Figure 4‑1, there are different persons with similar appearance, making this dataset quite challenging.

Besides, this dataset splits the bounding boxes into two categories: “good” and “junk” image. Intersection over Union (IoU) is an evaluation metric used to measure the accuracy of an object detector on a particular dataset. An image is marked as “good” (“junk”) if the IoU score between predicted boxes and ground truth is higher (lower) than 50% (see Figure 4‑2). Those “junk” image will not be taken into consideration when training and inferencing the model.

### DukeMTMC-reID Dataset

|  |
| --- |
|  |
| Figure 4‑3. The camera position setting for DukeMTMC-re-ID. |

DukeMTMT-reID [35] is the person re-identification subset of the DueMTMC dataset, which is a tracking dataset with 85 minute high-resolution video with manually labeled bounding boxes collected from 8 different cameras in Duke University. Besides, DukeMTMC-reID provides DukeMTMC-attribute with 23 types of attribute data annotation and DukeMTMC-pose for download. There are 8 static and non-overlapping cameras installed on the campus of Duke University, as shown in Figure 4‑3.

A total of 36411 images are sampled every 120 frames in videos and manually cropped to form 1404 people who were exposed to more than two cameras. Furthermore, the other 408 people are captured under only one camera to be the distractors. The whole dataset contains 16522 training images (702 identities), 2228 query images (another 702 identities) and a search library (gallery) of 17661 images which includes 408 distractors corresponding to junk images.

|  |
| --- |
|  |
| Figure 4‑4. The images distribution on DukeMTMC-reID training set. |

Figure 4‑4 shows the data distribution on DukeMTMC-reID training set. As we can see, the number of images from each identity is significantly different. The median of images per identity is 20 and some identities may contain lots of images (The 5388-th identity contains 426 images).

### Evaluation Metrics

Because the identities in the testing set do not exist in the training set, we cannot apply accuracy directly to evaluate the Re-ID algorithm. Instead, we create a sorting list by calculating the similarity of the query to all gallery images and the ordering of the list reflects the performance. In this thesis, Cumulative Match Characteristic (CMC) curve and the Mean Average Precision (mAP) is used to evaluate our method.

|  |
| --- |
|  |
| Figure 4‑5. The Cumulative Match Characteristic (CMC) curve. |

A CMC curve is used to assess the accuracy of algorithms that produce an ordered list of possible matches. For instance, the output of the algorithm would be a list of images from the training-set, ordered from most to least likely to be the test person. The larger area under the curve (AUC) the better the performance (see Figure 4‑5).

|  |
| --- |
|  |
| Figure 4‑6. Rank- calculation. |

The rank can be calculated as:

|  |  |
| --- | --- |
|  | 4‑1 |
|  | 4‑2 |
|  | 4‑3 |

Imagine a CMC curve where the rank 10 accuracy is 50%. This means that the correct match will occur somewhere in the top 10, 50% of the time. In general, the better the algorithm, the higher the rank- CMC-percentage.

As introduced in Section 2.4, the general definition for the Average Precision (AP) is the area under the precision-recall curve. The calculation of AP can be formulated as follows:

|  |  |
| --- | --- |
|  | 4‑4 |

where represents the number of the gallery images with the same identity to the query . denotes the number of all the gallery images. illustrates the precision at rank for the -th query. And is the identity of -th prediction. Mean AP (mAP) is calculated by averaging the AP of a set of queries, which is formulated as:

|  |  |
| --- | --- |
|  | 4‑5 |

where is the number of queries in the set and is the average precision for a given query . What the formula is essentially telling us is for a give query , we calculate its corresponding AP. Then the mean of all these AP scores would provide us with a single number called “mAP”, which quantifies how good our model is at performing the query. Besides, mAP is very sensitive to the ranking of retrieval result. In other words, the higher-ranked object contributes more to the average than the lower-ranked object does.

## Experimental Results

## Ablation Studies

1. REFERENCE

[1] L. Zheng, Y. Yang, and A. J. A. Hauptmann, "Person Re-identification: Past, Present and Future," vol. abs/1610.02984, 2016.

[2] A. Kläser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *BMVC*, 2008.

[3] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," vol. 60, no. 2 %J Int. J. Comput. Vision, pp. 91–110, 2004.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 9, pp. 1627-1645, 2010.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995/09/01 1995.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision,* vol. 88, no. 2, pp. 303-338, 2010/06/01 2010.

[7] H. Tang-Wei, Y. Yu-Huan, Y. Tso-Hsin, L. An-Sheng, F. Li-Chen, and Z. Yi-Chong, "Privacy free indoor action detection system using top-view depth camera based on key-poses," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016, pp. 004058-004063.

[8] T. Tseng, A. Liu, P. Hsiao, C. Huang, and L. Fu, "Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 4077-4082.

[9] S. Lin, A. Liu, T. Hsu, and L. Fu, "Representative Body Points on Top-View Depth Sequences for Daily Activity Recognition," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 2968-2973.

[10] R. B. Girshick, J. Donahue, T. Darrell, J. J. I. C. o. C. V. Malik, and P. Recognition, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," pp. 580-587, 2014.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886-893 vol. 1.

[12] Y. Zhang, C.-y. Wang, X. Wang, W. Zeng, W. J. a. C. V. Liu, and P. Recognition, "FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking," 2020.

[13] K. He, X. Zhang, S. Ren, J. J. I. C. o. C. V. Sun, and P. Recognition, "Deep Residual Learning for Image Recognition," pp. 770-778, 2016.

[14] H. Chen, B. Lagadec, and F. Bremond, "Learning Discriminative and Generalizable Representations by Spatial-Channel Partition for Person Re-Identification," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 2472-2481.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems,* vol. 25, 01/01 2012.

[16] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision,* vol. 115, no. 3, pp. 211-252, 2015/12/01 2015.

[17] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014*, Cham, 2014, pp. 818-833: Springer International Publishing.

[18] A. Bochkovskiy, C.-Y. Wang, and H. J. A. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," vol. abs/2004.10934, 2020.

[19] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in *NIPS*, 2014.

[20] C. Feichtenhofer, A. Pinz, A. J. I. C. o. C. V. Zisserman, and P. Recognition, "Convolutional Two-Stream Network Fusion for Video Action Recognition," pp. 1933-1941, 2016.

[21] K. Simonyan and A. J. C. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," vol. abs/1409.1556, 2015.

[22] C. Szegedy *et al.*, "Going deeper with convolutions," pp. 1-9, 2015.

[23] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks,* vol. 2, no. 5, pp. 359-366, 1989/01/01/ 1989.

[24] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 6, pp. 107-116, 04/01 1998.

[25] F. Schroff, D. Kalenichenko, J. J. I. C. o. C. V. Philbin, and P. Recognition, "FaceNet: A unified embedding for face recognition and clustering," pp. 815-823, 2015.

[26] A. Hermans, L. Beyer, and B. J. A. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," vol. abs/1703.07737, 2017.

[27] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," Cham, 2016, pp. 499-515: Springer International Publishing.

[28] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking Person Re-identification with k-Reciprocal Encoding," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3652-3661.

[29] G. E. Hinton, O. Vinyals, and J. J. A. Dean, "Distilling the Knowledge in a Neural Network," vol. abs/1503.02531, 2015.

[30] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net," in *ECCV*, 2018.

[31] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 34, 08/16 2017.

[32] X. Fan, W. Jiang, H. Luo, and M. J. J. V. C. I. R. Fei, "SphereReID: Deep Hypersphere Manifold Embedding for Person Re-Identification," vol. 60, pp. 51-58, 2019.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. J. I. C. o. C. V. Wojna, and P. Recognition, "Rethinking the Inception Architecture for Computer Vision," pp. 2818-2826, 2016.

[34] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-identification: A Benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116-1124.

[35] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. J. A. Tomasi, "Performance Measures and a Data Set for Multi-target, Multi-camera Tracking," vol. abs/1609.01775, 2016.