Global-local attention with triplet loss and label smoothed cross-entropy for person re-identification

Nha Tran, Toan Nguyen, Minh Nguyen, Khiet Luong, Tai Lam
Faculty of Information Technology, University of Education, Ho Chi Minh City, Viet Nam

ABSTRACT

Person re-identification (Person Re-ID) is a research direction on tracking and identifying people in surveillance camera systems with non-overlapping camera perspectives. Despite much research on this topic, there are still some practical problems that Person Re-ID has not yet solved, in reality, human objects can easily be obscured by obstructions such as other people, trees, luggage, umbrellas, signs, cars, motorbikes. In this paper, we propose a multi-branch deep learning network architecture. In which one branch is for the representation of global features and two branches are for the representation of local features. Dividing the input image into small parts and changing the number of parts between the two branches helps the model to represent the features better. In addition, we add an attention module to the ResNet50 backbone that enhances important human characteristics and eliminates irrelevant information. To improve robustness, the model is trained by combining triplet loss and label smoothing cross-entropy loss (LSCE). Experiments are carried out on datasets Market1501, and duke multi-target multi-camera (DukeMTMC) datasets, our method achieved 96.04% rank-1, 88.11% mean average precision (mAP) on the Market1501 dataset, and 88.78% rank-1, 78.6% mAP on the DukeMTMC dataset. This method achieves performance better than some state-of-the-art methods.

Keywords: Attention mechanism, Deep features, Multi-level feature, Multi-loss, Person re-identification

This is an open access article under the CC BY-SA license.

1. INTRODUCTION

Person re-identification is a research direction on tracking and identifying people in surveillance camera systems with non-overlapping camera perspectives. This is considered an important technology in the field of computer vision and has practical significance in building intelligent monitoring systems. It has many monitoring applications, such as crime detection, personal tracking, and activity analysis [1]–[3]. With the development of deep learning especially convolutional neural networks, person re-ID has made important progress in recent years.

In the past, researchers often used manual methods to solve the problem of a person re-identifying by extracting and selecting features [4]–[7]. Gray and Tao [8] proposed a partition strategy for extracting color and texture features by dividing the image of pedestrians into many horizontal stripes. Other researchers have also utilized more advanced partition methods. The pedestrian image was divided into several triangles to extract features by section in [9]. In [10], hue, saturation, value (HSV) histograms based on body parts, such as the head and trunk, were used to capture spatial information for positioning.

Although much research has been conducted on this topic, there are still some practical problems that have not been fully addressed in the field of person re-identification. One example is the scenario where the
subject to be re-identified is obscured by other objects, which has been studied by [11]. However, with the development of deep learning technology, researchers moved their focus to deep learning techniques, in which models automatically learn pedestrian features. These methods are based on deep learning and are classified into four main groups: Deep metric learning, local feature learning, generative adversarial networks, and Sequence feature learning.

For Deep metric learning method aims to determine the similarity or dissimilarity between two objects (in this case, a pedestrian) by learning distinguishing features from the input image and using a distance function. The distance between two people will have a small value if the same person is similar and vice versa. Deep metric learning is mainly used to constrain the learning process of discriminant features by designing a loss function for the model. Loss functions such as classification loss, verification loss, triplet loss, and contrastive loss are frequently used in research. In [12], triplet loss was used and achieved very good results (state-of-the-art) on both pre-trained and retrained models. The authors emphasized the importance of designing a good triplet loss and encouraged further exploration of the potential of this loss function.

For local feature learning method learns the distinguishing features to perform the classification of two objects as similar or different, these distinguishing features come from local features on the input image. Neural networks are used to automatically find local regions containing important information and then extract distinctive features from these regions. The local features in this method must be related to each other to ensure the accuracy of the output. Commonly used local feature learning methods are predefined stripe segmentation, multi-scale fusion, soft attention, pedestrian semantic extraction, and global-local feature learning [13]–[16].

A network called the part-based convolutional baseline (PCB) was proposed in [15]. The input image was divided into p parts evenly (p=6) and each classifier predicted the identity of the input image using Cross-Entropy loss as a supervision signal during training. During testing, either p pieces are concatenated to form the final descriptor of the input image. A strategy of learning discriminant features based on many details through the application of a global-local learning method was proposed in [17]. The author has designed the multiple granularity network (MGN) as a multi-branch deep learning network architecture in which one branch is for the representation of global features and two for the representation of local features. In the two branches of learning local features, the author uses the uniform partitioning method to divide the input image into many parts (horizontal division) and the diversity in the number of parts in different branches help the model represent the features in many details. This method has achieved a high performance and surpasses previous methods. In [18], the relationship between local features within an image and the relationship between features in different images was investigated. Local feature histograms were used and the attention mechanism was employed during training to assign varying levels of importance to different features.

For generative adversarial networks (GAN) [19] was introduced by Ian Goodfellow and developed rapidly in recent years. The main application of GAN is to generate new images for the purpose of expanding the dataset [20]. In person re-identification some researchers use GAN to generate new human images with differences in appearance, posture, even lighting contrast and resolution. The CycleGAN technique [21] has been utilized by some researchers to transform the style of images between different datasets. This inspired the proposal of PTGAN [22], a method that transforms the style of images in the domain of one dataset to the domain of another dataset while preserving the identity of the people in the original domain. The aim of this method is to convert the background and brightness of the original domain to the remaining domain in order to increase the diversity of the data. An adversarial network for hard triplet generation (HTG) was proposed in [23] in order to optimize the network's ability.

Sequence Feature Learning is used by some researchers to extract the important information contained in video sequences. This method takes a short video as input and uses both spatial and temporal cues to identify the object. Recognizing the lack of spatial and temporal constraints in many person re-identification methods, a method that enables the extraction of both spatial and temporal information was proposed in [24]. The main idea here is that when a person appears in the frame of one camera, for some time t (t is a small value) cannot appear in the frame of another camera. Due to temporal constraints, this method has helped to eliminate many misidentified human images.

In this study, we propose a multi-branch learning network, in order to represent global features and represent local features, the multi-branch model has been increasingly widely used in recent years [25]–[29]. Part-level feature learning is beneficial for learning the discriminatory re-ID model. Global feature learning learns representations over the whole image with no part constraints [30]. It is discriminatory when tracking someone who can pinpoint the exact location of the human body. When the human image is obscured, the learning part-level feature often achieves better performance by exploiting discriminatory body regions [14]. Because of its advantage in occlusion processing, we observe that most of the recently developed state-of-the-art methods adopt the model of feature aggregation, partial-level matching, combining feature-level matching and full human body features [28], [31].
Besides, designing the loss function is very important during training, the common approach is to use a combination of ID loss and metric loss. Our main contributions are:

- Multi-branch with convolutional block attention module (CBAM) [32] is attached to the backbone to help the model learn more detailed features. The features are then automatically grouped into sub-features, each of which helps to narrow the search space of the recognition target.
- Combination triple loss and smooth cross entropy loss to help the network learn more efficiently to improve performance.

2. METHOD

The proposed method is a network named GLML as shown in Figure 1 based on MGN [17] consisting of many branches including one branch for the representation of global features and two for the representation of local features. In the two branches of learning local features, using the homogenous partitioning method to divide the input image into several horizontal parts, and the diversity of the number of parts in different branches helps the model represent are featured in many details. The first part represents the global feature learning branch, which uses the input original image and feature extraction on it. The middle part (Branch 2) and the last part (Branch 3) represent two branches of learning local features, dividing the input image into many small parts and changing the number of parts between the two branches helps the model can represent better features. In addition, a CBAM attention module adds to the ResNet50 [33] backbone to enhance important human features and remove irrelevant information. To improve robustness, the model is trained by combining triplet loss and label smoothing cross-entropy loss (LSCE). The common loss function is constructed based on LSCE loss and triplicate loss.

![Figure 1. Multi-branch with CBAM and multi loss](image)

**Algorithm 1** The MLML Algorithm

**Input**: A fixed-size mini-batch consisting of P = 4 randomly selected identities and K = 4 randomly selected images per identity from the training set.

**Output**: Tensors of feature extraction and reID task.

**Initialization**: A fixed-size N of epochs.

**Repeat**

for each epoch i = 1 to N do

  for each iteration to max iterations do

    Backpropagate CNN and evaluate the loss L according to (1)

  end for

if i % 50 = 0:

  Start evaluate;
  Save the model with lowest loss L

end if

end for

**Until** maximum epochs (N = 800) reach.

---

*Global-local attention with triplet loss and label smoothed cross-entropy for person … (Nha Tran)*
To make a fast convergence of the training models, the Adam optimizer is adopted. The parameters of Adam have a weight decay of $5 \times 10^{-4}$, we set the initial learning rate to $2 \times 10^{-4}$, and decay the learning rate to $2 \times 10^{-5}$ and $2 \times 10^{-6}$ after training for 360 and 720 epochs. The total training process lasts for 800 epochs. To extract the discriminative features for person re-identification, ResNet-50 as the backbone network was used. The batch size of 16 was chosen based on the available hardware resources and the trade-off between training speed and stability. The Pytorch platform with T4 graphical processing unit (GPU) was adopted in our work. The images were resized to $384 \times 128$ and subjected to random erasing and horizontal flipping with a probability of 0.5 in order to augment the training data and improve the robustness of the model. We also followed the approach outlined in [34] to improve the performance of the model.

2.1. Attention module

The attention module helps the model learn and focus more on important information than on learning unhelpful background information. We added the CBAM block to the backbone, creating the model shown in Figure 2. the CBAM uses two attention modules: the channel attention module, as shown in Figure 3, and the spatial attention module shown in Figure 4. The envelope channel attention module consists of two feature maps, each consisting of two intermediate layers: average and maximum pooling. Both feature maps are combined by a shared multilayer perceptron layer (MLP), then the output of the feature map is added using the sigmoid activation function. Finally, the multiplicative features between the convolutional layer and the channel attention module were applied to the spatial attention module to determine the positions of the most important features in the given image.

![Figure 2. Convolutional block attention module architecture](image)

![Figure 3. Channel attention module](image)

![Figure 4. Spatial attention module](image)
2.2. Loss functions

To optimize the distance feature of each class in our model, we employed the triplet loss during training, while softmax cross-entropy loss was utilized to capture the classification difference in each class. The overall loss of our proposed method can be expressed as the sum of triplet loss and softmax cross-entropy loss, as shown in (1). However, most of the studies apply the batch-hard triplet loss [12], [35], an improved version based on the original semi-hard triplet loss, as shown in (2).

\[ L_g = L_{\text{triplet}} + L_{\text{sce}} \]  
\[ L_{\text{triplet}} = \sum_{k=1}^{N_k} \sum_{m=1}^{N_M} [\alpha + \max_{n=1,...,M} q_{k,m}^A - q_{k,n}^P - \min_{n=1,...,N} \| q_{k,m}^A - q_{k,n}^N \| 2]^+ \]  

In our model, for the purpose of classifier learning, we have adopted the softmax loss function. This loss function is widely used for multiclass classification problems and is mathematically expressed as the negative logarithm of the softmax function. Specifically, the softmax loss function computes the difference between the predicted probability distribution and the true probability distribution of the classes, as shown in (3).

\[ L_{\text{sce}} = -\sum_{i=1}^{N} \log \frac{e^{h_i y_i}}{\sum_{k=1}^{N} e^{h_k f_i + b_k}} \]  

To enhance the robustness of our model and prevent overfitting, we employed the technique of Label Smoothing [36]. This method modifies the target distribution by allocating some probability mass to non-target classes during training, as defined by (4). The soft-margin \( \varepsilon \) is used to reduce model overconfidence. In our experiments, we set \( \varepsilon \) to 0.2.

\[ L_{\text{sce}} = \sum_{i=1}^{N} -q_i \log (p_{i}) \begin{cases} 
q_i = 0, & y \neq i \\
q_i = 1, & y = i 
\end{cases} 
q_i = \begin{cases} 
1 - \frac{N-1}{N} & \varepsilon if i = y \\
\varepsilon / N & otherwise 
\end{cases} \]  

3. RESULTS AND DISCUSSION

3.1. Datasets

**Market-1501** [37]: This large data set was published in 2015, Market-1501 was collected by 5 high-resolution cameras and 1 low-resolution camera in front of a supermarket at Tsinghua University, China. The dataset contains 1501 different people with a total of 32668 images. Compared to CUHK03 [38], Market-1501 has more images and contains many confounding factors (images of people are obscured, photos only show part of the body, ...), so this dataset is evaluated as closer to reality than CUHK03.

**DukeMTMC-reID** [39]: dataset collected at Duke University, USA through 8 HD still cameras. DukeMTMC-reID includes a training set containing 16522 images of 702 different people, a query set of 2228 images of 702 other people (different from the training set), and a gallery of 17661 images. Both datasets include bounding box annotations and additional metadata such as camera IDs and timestamps and are commonly used to evaluate the performance of person re-identification algorithms using standard evaluation metrics such as rank-1 accuracy and mean average precision (mAP). The Market1501 and DukeMTMC datasets have been widely used in a variety of person re-identification tasks, including cross-camera person search, video-based person re-identification, multi-camera tracking, and pedestrian attribute recognition.

3.2. Result and discussion

Table 1 demonstrates that our model obtained the highest accuracy rates of 87.77% and 88.11% on the Market-1051 dataset. For the DukeMTMC-reID dataset, we achieved a mAP of 76.97% and 78.60%. These results indicate that utilizing an attention mechanism enhances significant human features while filtering out irrelevant information. Furthermore, combined triplets’ loss, and smooth label softmax cross-entropy loss help the network to learn more effectively.

In order to gain a better understanding of the results obtained by our proposed model. We have compared it with several other methods using two datasets: Market – 1051 and DukeMTMC-reID. Table 2 and Table 3 shows that our proposed method has achieved comparable performance to some of the state-of-the-art methods developed in recent years.
REFERENCES


Global-local attention with triplet loss and label smoothed cross-entropy for person … (Nha Tran)


BIographies of authors

Nha Tran received BS. Computer Science (2014), MS. Computer Science (2019) from HCM University of Education. He is currently a lecturer of Information Technology Faculty, Ho Chi Minh city University of Education. His research interest includes computer vision, information retrieval, affective computing, machine learning. You can contact him via email: nhatt@hcmue.edu.vn.

Toan Nguyen received his B.Sc. (Computer Science) from Ho Chi Minh City University of Education in 2022. His research includes machine learning, deep learning, computer vision, image processing. He can be contacted at email: toannn20@gmail.com.
Minh Nguyen He is currently a 3rd-year student at the Ho Chi Minh City University of Education, Ho Chi Minh City, Vietnam. His research includes machine learning, deep learning, and computer vision. He can be contacted at email: nguyendatminhvn@gmail.com.

Khiet Luong received BS. Information Technology (2016), MS. Computer Science (2018). He is currently a lecturer at the University of Education, Ho Chi Minh City, Vietnam. Concurrently, his research interests include computer vision, Software engineering, and machine learning. You can contact him via email: khietltn@hcmue.edu.vn.

Tai Lam received his B.Sc. (Computer Science) from Ho Chi Minh City University of Education in 2022. He is currently an AI engineer at Emage Development company. His research includes machine learning, deep learning, computer vision. You can contact him via email: lamtai2105@gmail.com.