DEEP JOINT DISCRIMINATIVE LEARNING FOR VEHICLE RE-IDENTIFICATION AND RETRIEVAL

Yuqi Li, Yanghao Li, Hongfei Yan, Jiaying Liu*

Peking University, Beijing, P.R.China, 100871

ABSTRACT

In this paper, we propose a novel vehicle re-identification method based on a *Deep Joint Discriminative Learning* (DJDL) model, which utilizes a deep convolutional network to effectively extract discriminative representations for vehicle images. To exploit properties and relationship among samples in different views, we design a unified framework to combine several different tasks efficiently, including identification, attribute recognition, verification and triplet tasks. The whole network is optimized jointly via a specific batch composition design. Extensive experiments are conducted on a large-scale VehicleID [1] dataset. Experimental results demonstrate the effectiveness of our method and show that it achieves the state-of-the-art performance on both vehicle re-identification and retrieval.

Index Terms— Joint Discriminative Learning, Vehicle Re-Identification, Vehicle Retrieval

1. INTRODUCTION

Vehicle search and re-identification is an important problem in computer vision, which has many practical applications like video surveillance systems. Although the license plate provides a unique ID for a vehicle, sometimes it is still not easy to recognize its plate. For example, the resolution of images is not enough due to the environment or the camera, or the plate is occluded or removed. Thus, vehicle re-identification based on appearance information still plays an important role for real applications.

Although vehicle identification problem is of a great importance, most previous object identification works focus on human face or person [2, 3, 4, 5]. However, their targets are similar, which are to learn discriminative representations for images. Recently, deep convolutional network has also demonstrated its great power in identification tasks. In [2], Yi *et al.* introduced a deep network to directly classify all identities (about 10,000 classes) for face recognition. Then, a pair-wise verification loss [3] is proposed to be combined with identification loss to help reduce intra-class variations by pulling features of same identity together. Similar verification loss is also utilized in person identification [6, 7]. Another successful deep learning framework is triplet loss for both face recognition [5] and person re-identification [8]. It learns the embedding representations in the deep convolution network by optimizing the triplet loss, which is under the assumption that samples of the same identity should be closer from each other than samples of different identities.

Different with face or person identification problem, vehicle identification could be more challenging since it is really hard to discriminate vehicles with similar visual appearance which belongs to the same model. Most previous related works about vehicle focus on the vehicle model classification [9, 10] which only recognize vehicle models instead of further identities. Recently, Liu *et al.* [1] presented a new large-scale vehicle re-identification database 'VehicleID', which is collected from the real surveillance cameras and labeled in identity level. The large scale of the dataset facilitates the recent deep learning models, which have been proved more effective and robust for many vision tasks, to apply to the vehicle identification problem. Inspired by some state-of-the-art methods in face recognition [3, 5], in this paper, we propose a Deep Joint Discriminative Learning (DJDL) model for vehicle re-identification and retrieval problem.

The proposed DJDL model is an end-to-end multi-task deep framework, which aims to learn a deep convolutional model that can extract discriminative features for vehicle images. The overall network is illustrated in Fig. 1. DJDL incorporates four different subnetworks in a unified framework to capture different properties and relationships among samples. The four subnetworks includes identification, attribute recognition, verification and triplet tasks. Identification and attribute recognition focus on the individual samples to exploit their own specific properties. The idea of verification task is to constrain relationship between two samples (*eg.* minimize distance between two samples of same identity) while triplet task is responsible for constraining the relative distance among three samples. At the same time, we propose an efficient batch composition design to jointly optimize the four objective functions. Experiments on vehicle re-identification and retrieval demonstrate the complementary effect among the four tasks. The results also show that our DJDL model achieves promising results and outper-

∗ Corresponding author

This work is supported by 973 with Grant No.2014CB340400, NSFC with Grant No.U1536201 and NSFC with Grant No.61472013.

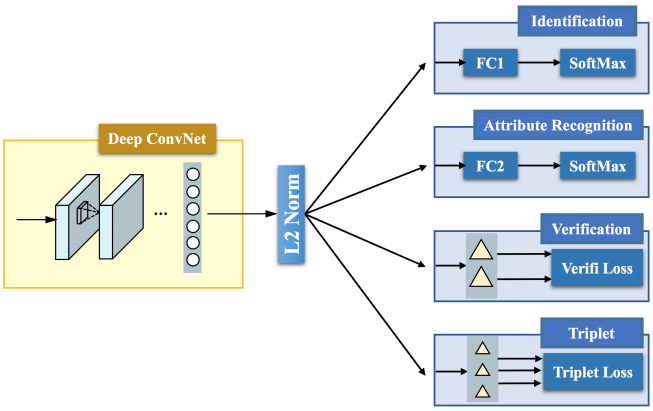


Fig. 1. Architecture of the proposed Deep Joint Discriminative Learning framework for Vehicle Re-Identification.

forms other state-of-the-art approaches.

The reminder of the paper is organized as follows: we discuss about the details of our proposed Deep Joint Discriminative Learning method in Section 2, and present the experimental results in Section 3. The conclusion is drawn in Section 4.

2. DEEP JOINT DISCRIMINATIVE LEARNING

In this section, we illustrate the detains of our proposed Deep Joint Discriminative Learning model. Specifically, we first introduce the overall network structure, and then introduce multiple tasks respectively. At last, we explain some training and optimization details for our proposed network.

2.1. Network Architecture

The overall network architecture of the proposed DJDL model is illustrated in Fig. 1. It is essentially a multi-task joint learning network that joint optimizes several different objectives and aims to learn a deep representation of vehicle appearance that is discriminative for different vehicle identities. The network consists of five parts: a base network shared by different branches to extract representative features for each image and four subnetworks of different tasks, including identification, attribute recognition, verification and triplet subnetworks. The base network can be a common deep convolutional network such as Inception-BN [11], VGG [12] or ResNet [13], which is pre-trained on the ImageNet. For classification tasks (identification and attribute recognition), a single vehicle image individually is fed into the network. Two images are fed for verification subnetwork while three images for triplet subnetwork. Note that different image inputs are shared with the same base convolutional network for feature extraction. The different tasks are jointly optimized in the unified network at the same time.

2.2. Identification and Attribute Recognition Losses

Identification subnetwork considers each input image individually to predict its identity label. As the Fig. 1 shows, after the deep feature f_i extracted by the *L2 Normalization* layer, we add a fully connected layer ($m \times n$-dim) for identification classification, where m is the feature dimension and n is the number of training identities in the dataset. As conventional recognition approaches, we use the softmax loss for the identification subnetwork:

$$L_{identi}(f_i) = -\sum_{j=1}^{n} p_j \log \hat{p_j}, \tag{1}$$

where $\hat{p_j}$ is predicted probability and p_j is the target probability ($p_j = 1$ for $j = v_i$ otherwise $p_j = 0$ where v_i is the identity label of the input image).

Since the appearance information is the key clue for vehicle identification and verification, we utilize the vehicle attribute information in the dataset to further improve the performance. In VehicleID [1] dataset, images are labeled by some attributes, such as color and vehicle model information. Thus we propose an attribute recognition subnetwork to explicitly learn to recognize these attributes. Specifically, we also use the standard softmax loss for the attribute recognition subnetwork:

$$L_{attri}(f_i) = -\sum_{k=1}^{n_{attri}} \sum_{j=1}^{n_k} a_j^k \log \hat{a_j^k}, \tag{2}$$

where n_{attri} is the number of attributes, n_k is the number of labels for k-th attribute, $\hat{a_j^k}$ is the predicted probability for k-th attribute and a_j^k is the corresponding target probability.

2.3. Verification and Triplet Losses

To further boost the efficiency of the learned deep representation, we combine the identification and attribute recognition losses with two discriminative objectives, including the verification loss and the triplet loss.

The verification subnetwork is a pair-wise siamese network, which first takes two feature vectors f_i and f_j as input, and then calculates the similarity according to the label of two images. After the normalization, we directly use Euclidean Distance to measure the similarity of the two input images. Specifically, the distance should be small if the two images are belong to the same identity where the distance should be large for two images with different identities. We adopt a distance function similar to [3] and the loss function can be formulated as:

$$L_{verif}(f_i, f_j) = \begin{cases} \frac{1}{2}\|f_i - f_j\|_2^2, & v_i == v_j, \\ \frac{1}{2} max(0, \alpha - \|f_i - f_j\|_2)^2, & v_i \neq v_j, \end{cases} \tag{3}$$

where α is the margin parameter which enforces the distance of different identities larger than α.

The triplet subnetwork involves three samples each time, which considers the relative distance among identities while verification subnetwork constrains the absolute distance between two identities. The triplet loss [5] wants to ensure that an *anchor* image i is closer to all *positive* images j of the same identity ($v_i = v_j$) than it is to other *negative* images k of different identities ($v_i \neq v_k$). Specifically, the triplet loss can be formulated as:

$$L_{triplet}(f_i, f_j, f_k) = max(0, \|f_i - f_j\|_2^2 - \|f_i - f_k\|_2^2 + \beta), \tag{4}$$

where β is a margin parameter which enforces the distance between positive pairs and negative pars larger than β.

2.4. Training and Optimization

The above four type of tasks are optimized at the same time in a unified network. Thus, we can obtain the total objective function:

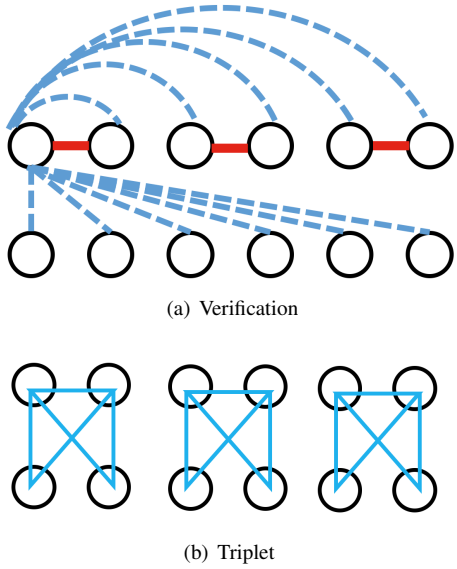
$$L = L_{identi} + L_{attri} + L_{verif} + L_{triplet}. \tag{5}$$

We use the conventional Stochastic Gradient Descent (SGD) optimization in the training. Thus one natural problem is how to jointly optimize four losses together for a single batch in the training, especially for verification and triplet tasks. For identification and attribute recognition tasks, there is no special treatment since each sample in a batch is treated individually. Thus, we adopt a specific batch composition design during the batch training.

In the implementation, we ensure that half of samples in a batch are generated by positive pairs while the other half are selected randomly. Fig. 2 illustrates an example of one generated batch for verification and triplet, respectively. For verification loss, every two samples in a batch is treated as a verification sample. Fig. 2(a) shows three positives pairs and some negative pairs which connected to the first sample. Since the first half samples are composed of positive pairs, it ensures there are positive verification samples for each image in the first half of a batch. For triplet loss, the anchor and the positive ones are selected from the first half of the batch while the negative ones are selected from the second half of the batch. Fig. 2(b) shows six group of triplets in a batch. During training, each batch is randomly generated but meets the above constraint.

3. EXPERIMENTS

In this section, we evaluate our DJDL method on the vehicle re-identification and retrieval tasks. We adopt the recent released large-scale vehicle dataset VehicleID [1], which has 221,763 images of 26,267 vehicles in total. Follow the protocols in [1], we use the three testing sets (*i.e.*, small, medium and large) with different size for the vehicle retrieval and re-identification tasks.



(a) Verification

(b) Triplet

Fig. 2. An example of a training batch. The top samples are generate by positive pairs (with same identity) while the bottom samples are selected randomly. (a) Training pairs for verification task. Red lines corresponds to positive pairs while blue lines corresponds to negative pairs (only negative pairs connected to the first sample are shown). (b) Training triplets for triplet task. Each triangle corresponds to one triplet sample where the top two samples correspond to anchor and positive samples while the bottom sample is the negative one. This figure is best viewed in color.

3.1. Implementation Details.

We use MXNet [14] package in the implementation. We adopt Inception-BN [11] as the base convolutional network in the experiments. During training, the input image is randomly cropped at 224×224 from the resized image (the short edge is resized as 256) and randomly mirrored horizontally. The training images are shuffled and generated corresponding batches as explained in Sec. 2.4. The batch size is set as 64. The initial learning rate starts from 0.01 and is divided by 10 at 50, 75 epochs for total 100 epochs. For the four different objective functions in our network, we simply assign 1 as the gradient weight for the losses. The margin parameters α and β are set as 0.9.

For attribute recognition task, we only use the vehicle model labels in the VehicleID dataset [1] since the color labels is incomplete for all vehicles images. For vehicle retrieval and re-identification tasks, we first extract the normalized feature representations for images in both gallery and probe sets, the similarity between arbitrary two vehicle images is measured by the L2 distance. To accelerate the retrieval process, we

use the fast approximation nearest neighbor searching library Flann [15].

3.2. Vehicle Retrieval

We first evaluate our method on the vehicle retrieval task following the widely used protocol in retrieval task, mean average precision (MAP). We follow the split strategy in [1] for the three testing image sets with different sizes.

At first, we adopt an ablation experiment about different components of our method. Table 1 illustrates the results of different methods. It shows that the performance improves consistently when incorporating identification, attribute recognition, verification and triplet tasks respectively. The final model *Identi+Attri+Verifi+Triplet* achieves the best performance, which demonstrates the effectiveness of multi-task design.

Method	Small	Medium	Large
Identi	0.712	0.684	0.670
Identi+Attri	0.718	0.686	0.672
Identi+Attri+Verifi	0.731	0.705	0.689
Identi+Attri+Verifi+Triplet	**0.786**	**0.747**	**0.720**

Table 1. MAP of Different DJDL models of Vehicle Retrieval Task.

Table 2 shows the final retrieval results compared with other state-of-the-art methods. Our DJDL method significantly outperforms the other methods, including Mixed Diff+CCL [1] and HDC + Contrastive [16], on all three testing datasets.

Method	Small	Medium	Large
VGG+CCL [1]	0.492	0.448	0.386
Mixed Diff+CCL [1]	0.546	0.481	0.455
HDC + Contrastive [16]	0.655	0.631	0.575
Identi+Attri+Verifi+Triplet	**0.786**	**0.747**	**0.720**

Table 2. MAP of Vehicle Retrieval Task.

3.3. Vehicle Re-Identification

In this section, we evaluate the DJDL method on the vehicle re-identification task. We follow the same evaluation protocols and train/test split settings in [1]. Fig. 3 illustrates the CMC curve, which is a common evaluation metric for re-identification problem, on the small size testing dataset. Table 3 illustrates the results on Top 1 and Top 5 match rate compared with other methods. From the results, we can see that our DJDL method achieves superior performance over other state-of-the-art methods on all three testing datasets. At the same time, after incorporating different tasks together, the match rate further increases consistently especially when combining all four tasks, which reveals the significant advantages of our Joint Discriminative Learning framework.

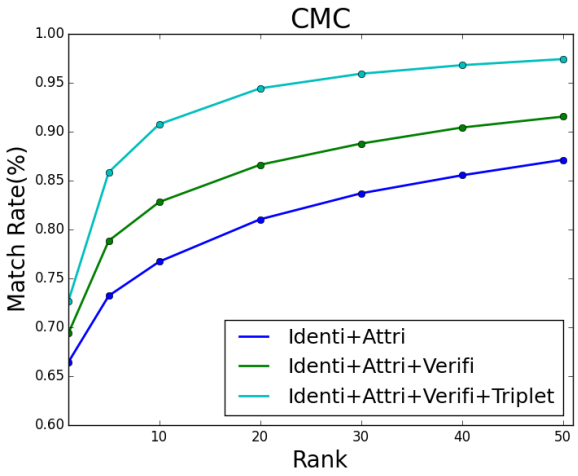


Fig. 3. CMC on VehicleID Dataset (Gallery size=800).

Method	Protocol	Small	Medium	Large
VGG+CCL [1]		0.436	0.370	0.329
Mixed Diff+CCL [1]		0.490	0.428	0.382
Identi+Attr	Top 1	0.670	0.667	0.651
Identi+Attr+Verifi		0.689	0.687	0.661
Identi+Attr+Verifi+Triplet		**0.723**	**0.708**	**0.680**
VGG+CCL [1]		0.642	0.571	0.533
Mixed Diff+CCL [1]		0.735	0.668	0.616
Identi+Attr	Top 5	0.735	0.729	0.716
Identi+Attr+Verifi		0.781	0.765	0.737
Identi+Attr+Verifi+Triplet		**0.857**	**0.818**	**0.789**

Table 3. Match Rate (Top 1 and Top 5) of Vehicle ReID Task.

4. CONCLUSION

In this paper, a novel Deep Joint Discriminative Learning (DJDL) model is proposed for vehicle re-identification and retrieval problem. We exploit a unified deep learning framework which incorporates four different type of tasks. The four tasks benefit each other due to their different properties and help to learn a deep convolutional network that could extract discriminative representations for vehicle images. During training, we jointly optimize the whole network by a specific designed batch composition scheme. Experimental results on a large-scale vehicle dataset demonstrate the effectiveness and necessity of each component of DJDL model. The results reveal that DJDL achieves superior results over several state-of-the-art approaches on both re-identification and retrieval tasks.

5. REFERENCES

[1] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *CVPR*, 2016.

[2] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, 2014.

[3] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," in *NIPS*, 2014.

[4] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deeply learned face representations are sparse, selective, and robust," in *CVPR*, 2015.

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[6] Rahul Rama Varior, Mrinal Haloi, and Gang Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*, 2016.

[7] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian, "Deep transfer learning for person re-identification," *arXiv preprint arXiv:1611.05244*, 2016.

[8] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[9] Yen-Liang Lin, Vlad I Morariu, Winston H Hsu, and Larry S Davis, "Jointly optimizing 3d model fitting and fine-grained classification.," in *ECCV*, 2014.

[10] Edward Hsiao, Sudipta N Sinha, Krishnan Ramnath, Simon Baker, Larry Zitnick, and Richard Szeliski, "Car make and model recognition using 3d curve alignment," in *WACV*, 2014.

[11] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[14] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang, "MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems," *NIPS Workshop on Machine Learning Systems*, 2016.

[15] Marius Muja and David G Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *VISAPP*, 2009.

[16] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang, "Hard-aware deeply cascaded embedding," *arXiv preprint arXiv:1611.05720*, 2016.