Joint Attribute Learning for Deep Person Re-Identification

Jayant Agrawal
Department of Computer Science and Engineering
IIT Kanpur
Email: agjayant@cse.iitk.ac.in

Vinay Namboodiri
Department of Computer Science and Engineering
IIT Kanpur
Email: vinaypn@cse.iitk.ac.in

***Abstract*—Person Re-Identification(Re-ID) is the task of identifying the queried person from images, taken from different non-overlapping cameras. Like many other problems in computer vision, large datasets and the rise of deep learning methods has brought about great development in Re-ID, also. In this work, we propose a deep multi-task CNN, Sharenet, to jointly learn mid-level human attributes and feature embeddings for re-id, in an end-to-end setup. We start by looking at the triplet networks and then show that jointly learning human attributes in a multi-task setup improves the quality of the learned feature embedding, which leads to improved Re-ID performance. We report competitive Re-ID results compared with the state-of-the-art methods on the dataset.**

I. Introduction

The aim of this work is to improve the quality of the learned feature embedding for a pedestrian image. An improved feature embedding can further result in a better performance at Person Re-ID. Person Re-Identification is the task of identifying a person given a query image, from a dataset of images taken from different non-related cameras.

We first start by looking at the Triplet Networks in Section II, and how they can be used for Re-ID. Triplet Networks have been used extensively for ranking and retrieval tasks, such as in [3] and [7]. In the recent years, they have been used for Re-ID also.

We look at multi-task Networks in Section III-A as a way to improve the quality of the feature embedding. The idea is to choose a task which is closely related to the task of feature extraction and can provide useful extra supervision. Multi-Task Learning is also a common feature in the recent works on Person Re-ID.

The task of attribute detection is chosen as a source of extra supervision. We look at the relationship between attributes and features and conduct experiments to get the proper architecture that can best use the supervision from attributes.

Finally, we conduct experiments to show that attributes actually help the performance in re-id and can give results which are competitive compared to the existing state-of-the-art.

II. Triplet Networks

The task of Person Re-ID involves finding the best match for a given query image of a pedestrian. We, therefore, need a model that gives us an efficient and an effective way to say whether a gallery image has the same person as that in the query image or not. The model should be able to learn a feature space, where the images are projected in such a way that the images having the same person are closer and those with different persons are 'comfortably' far away, with respect to some distance metric. By 'comfortably', we mean that there should be a fair *margin*. Triplet Loss[15] is one such loss function with precisely, the same aim.

$$\mathcal{L}_{triplet}(f) = \sum_{a,p,n} m + D(f(a), f(p)) - D(f(a), f(n))$$

where a,p,n denote anchor image, positive image and negative image respectively. Positive image has the same person/identity as that of the anchor image, while a negative image has a different person. f is any feature embedding function.

$\mathcal{L}_{triplet}$ encourages the positive image to be closer to the anchor image than the negative image, with respect to f. Formally, it enforces the following:

$$D(f(a), f(p)) < D(f(a), f(n))$$

m is simply the margin parameter, which says that the difference between $D(f(a), f(p))$ and $D(f(a), f(n))$ should be atleast m.

A deep triplet network (Figure 1) is a network with three branches with shared parameters in all the layers. Anchor, positive and negative images are passed through the three layers respectively. The features from the last layer are fed into the the triplet loss layer.

III. Proposed Approach

Triplet networks are known to perform well for ranking tasks. It has been shown to be very effective for Person Re-Id in the past, also. Our goal is to find an effective way to further improve the quality of the learned feature embedding. We show that multi-task learning is an effective way to do so.

A. Multi-Task Learning

In the past few years, multi-task learning has proven to be successful for recognition tasks such as classification,

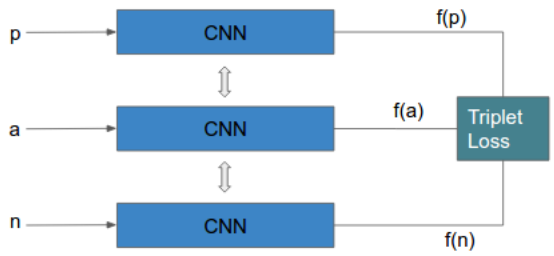

Fig. 1: Triplet Network

detection or localisation. The idea is to jointly train more than one task, each having it's own loss function, such that each task benefits from the supervision provided by the other taksk.

It is important to select the tasks wisely, such that they are co-related enough to be able to feed-off each other's supervision. The choice of architecture and the amount of sharing(if any) of information is also crucial.

For the task of Person Re-Identification, we choose a task which provides supervision which can help improve the learned feature embedding. The task that we chose, is detection of human attributes.

B. Human Attributes

For our problem, we focus on the mid-level human attributes(such as color of upper body clothes, lower body clothes,length of hair etc). The problem of human attributes detection has been mostly seen as an instance of multi-label learning.

Attributes have been used for re-id in the past with limited success. In [8], the last layer of a convnet is used as attributes and the same vector is used as the feature embedding for evaluating re-id.

C. Attributes and Features

Learning a discriminative feature embedding and attribute detection are closely related tasks. An attribute is more or less a combination of some set of features from a certain region in an image.

The lower level features learned in the lower layers of a deep neural net is bound to lead to a 'good' feature embedding if it performs well for the task of attribute detection. Therefore, it makes sense to share the weights in the lower layers of a convolutional neural network and use a multi-task loss for training.

The more closely related the tasks are, more is the amount of information they should share. We experiment with different architectures(shared information varies) and establish that the tasks of attribute detection and feature extraction are indeed very closely related in Section IV-A.

D. Sharenet

Using the above inferences, we propose a model, where the lower layers of a cnn are shared, which then branches into separate last layers for the the two tasks. The last layer in the attribute branch has k nodes, for a dataset with k attributes. The ground truth attribute vector is a binary vector of size k, where each element is 1 if the attribute is present in the image, otherwise 0. Sigmoid Cross Entropy Loss has been used as the Loss function.

The feature branch uses the Triplet Loss Function, which extracts triplets, $t = \{a, p, n\}$, from a mini-batch. The negative samples are extracted using hard-negative mining [7] within the mini-batch in one iteration.

The multi-taks loss, $\mathcal{L}_{mt}$, for sharenet is thus,

$$\mathcal{L}_{mt}(f, \mathcal{B}) = \mathcal{L}_{triplet}(f, \mathcal{B}) + \alpha * \mathcal{L}_{atr}(\mathcal{B})$$

where $\mathcal{B}$ is the mini-batch. $\mathcal{L}_{triplet}(f, \mathcal{B})$ gives the sum of the Triplet loss for all triplets in $\mathcal{B}$ and $\mathcal{L}_{atr}(\mathcal{B})$ is the sum of cross-entropy loss for all images in $\mathcal{B}$. α is the loss-weight parameter. The loss function is thus a weighted sum of the triplet loss and the cross-entropy loss.

Sharenet-5 which shares five lower layers is shown in Figure 2. A question that arises is the number of layers to be shared between the tasks. We analyse this in Section IV-A.

IV. Experimental Results

A. Number of Shared Layers

The number of shared layers in a multi-task setup depends on the combination of the tasks involved. Strongly co-related tasks may need more amount of sharing. Weakly co-related tasks may perform badly when more-than-required layers are shared. Different amount of sharing leads to different type of architectures. There are a lot of combinations possible. We look at a few of those.

For this experiment, we chose a small person tracking dataset, TownCentre. The architectures considered were Sharenet-5, Sharenet-6 and Sharenet-7, with 5, 6 and 7 shared layers respectively.

From Figure 3, one can observe that Sharenet-7 gives the best results. This shows that attribute detection and feature extraction are indeed very closesly related tasks for pedestrian images.

B. Dataset

We chose Market-1501 [2], as the dataset for our experiments. It is one of the largest Person Re-ID Datasets and the goto choice for benchmarking deep re-id models. It has $\tilde{3}$3k gallery images and $\tilde{3}$.5k query images taken from 6 cameras. Training set has images of 751 identities and the testing set has 750 identities. 500k distractors in the test set makes it even more challenging.

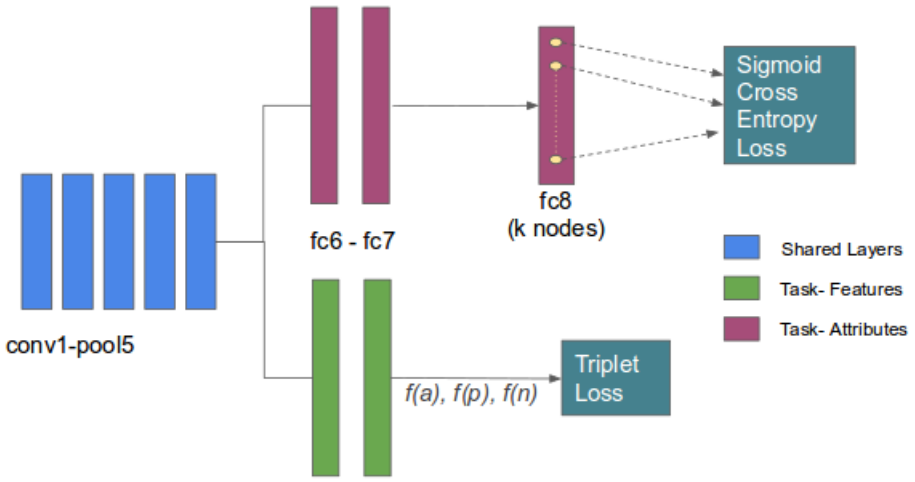

Fig. 2: *Proposed Model:* ShareNet-5, Base-*AlexNet*

Architecture	Rank-1
Triplet	61.3 %
Sharenet-5	73.15 %
Sharenet-6	74.52 %
Sharenet-7	82 %

Fig. 3: Rank-1 Results for **TownCentre**, Sharenet-*i* : *i* shared layers

Architecture	**Rank-1 (Single Query)**
LOMO [4]	26.07%
BoW [2]	34.68%
DADM [8]	39.4%
SiameseA[1]	41.24 %
SiameseR[1]	60.12%
IdentificationA[1]	56.03%
Baseline1-TripletA	58.31 %
Sharenet-7A(Ours)	**60.47 %**

Fig. 4: Rank-1 Results for **Market-1501**, Sharenet-*i* : *i* shared layers, A:*AlexNet* [10], R: *ResNet-50*[9]

C. Implementation Details

All the experiments were conducted on Titan X, with 12 GB of RAM. We experimented with the *Alexnet* [10] and the *Resnet* [9] Architectures.

For *Alexnet*-Baseline, batch size was kept to 100. Training was started with a learning rate of 0.001 and stepped down by a scale of 10 after regular intervals. For Sharenet-7 with alexnet as the base, α was set to 0.01 and a fully connected layer with size 35 was inserted for attributes in the dataset[14].

For *Resnet*-Baseline, batch size was kept to 32, due to the size of the model. The last layer was chopped off and a fully connected layer of size 1024 followed by batch normalization and relu layer and another fully connected layer with 751 nodes was inserted. Training was started with a learning rate of 0.0001 and stepped down by a scale of 10 after regular intervals. For Sharenet-7 with resnet as the base, α was set to 0.05 and a fully connected layer with size 35 was inserted in a new branch for attributes in the dataset[14].

D. Quantitative Results

Comparison with the state-of-the-art methods is shown in Figure 4. It can be seen that our model with alexnet as the base performs better than all the other architectures with alexnet as the base.

V. Conclusion

Inferring from the above experiments, we show that multi-task learning can be applied to ranking and retrieval problems like person re-identification, also. Jointly learning Attributes along with features helps improve the quality of features which further improves the performance in Person Re-Identification.

We show that feature extraction and attribute detection are strongly co-related tasks and that sharing of weights can be done up-to the semantic level. This gives us useful insights for the tasks of attributes and feature extraction.

VI. Future Work

During the course of this project, we realised that attribute detection is more closely related to the task of object detection, rather than image classification. Whether a person is wearing 'shorts' or 'pants' does not have any relation with his/her 'hair color'.

In this work, for classifying whether an attribute is present in an image or not, features from the entire image are considered. A more region based approach/model needs to be adopted for

Fig. 5: Qualitative Results

better attribute extraction.

Jointly learning this region-based model with a feature embedding may further lead to improvements in the results of Person Re-Identification.

REFERENCES

[1] Liang Zheng, Yi Yang, and Alexander G. Hauptmann Person Re-identification: Past, Present and Future arXiv 1610.02984, 10 Oct 2016

[2] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, Qi Tian Scalable Person Re-identification: A Benchmark 2015 IEEE International Conference on Computer Vision

[3] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet:A Unified Embedding for Face Recognition and Clustering. In CVPR, 2015.

[4] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 21972206.

[5] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints in IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 22882295.

[6] A. Mignon and F. Jurie, Pcca: A new approach for distance learning from sparse pairwise constraints, in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 26662672

[7] Xiaolong Wang and Abhinav Gupta Unsupervised Learning of Visual Representations using Videos Proc. of IEEE International Conference on Computer Vision (ICCV), 2015

[8] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. arXiv:1605.03259, 2016

[9] K. He, X. Zhang, S. Ren, and J. Sun Deep residual learning for image recognition in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton Imagenet classification with deep convolutional neural networks in Advances in Neural Information Processing Systems, 2012, pp. 10971105

[11] Z. Zheng, L. Zheng, and Y. Yang. A Discriminatively Learned CNN Embedding for Person Re-identification ,arXiv preprint arXiv:1611.05666, 2016

[12] Z. Zheng, L. Zheng, and Y. Yang Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro arXiv preprint arXiv:1701.07717, 2017

[13] S. Chopra, R. Hadsell, and Y. LeCun Learning a similarity metric discriminatively, with application to face verification. In CVPR, 2005

[14] lin, Yutian and Zheng, Liang and Zheng, Zhedong and, Wu Yu and, Yang, Yi ,Improving Person Re-identification by Attribute and Identity Learning, arXiv preprint arXiv:1703.07220 , 2017

[15] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In CVPR, 2015