Attributes classification and Person Re-Identification

Samuel Bortolin
University of Trento
samuel.bortolin@studenti.unitn.it

Alessandro Grassi
University of Trento
alessandro.grassi@studenti.unitn.it

Davide Lusuardi
University of Trento
davide.lusuardi@studenti.unitn.it

Abstract

Person re-identification and attribute classification are two related tasks that try to learn pedestrian descriptors. Attributes contain detailed local descriptors of pedestrians and are beneficial in allowing the re-identification model to learn more discriminative feature representations [1]. In this work, we propose an attribute and person recognition network able to detect with a good accuracy the attributes of a person in an image and also able to learn a re-identification embedding. The experimental results on Market-1501 dataset [2] demonstrate that by learning a more discriminative representation, the network achieves good re-identification performance on unseen pedestrian identities.

1. Introduction

Image classification is one of the most important research areas in the deep learning community. Its purpose is to find a way to recognize the subject of an image. Although it seems a specific task, it finds a variety of possible applications that include object detection and image segmentation. The reason of that is because the weights the network learns are not just specific for the classification purposes but are capable of generalize the content of the picture and produce useful and meaningful features. In our work we highlight this performing person re-identification using the same network that performed the classification task. We decided to use a pretrained ResNet-50[3] as the backbone feature extractor and a set of linear classifiers to classify each attribute. To make the network resistant to noise, we performed a series of transformations on the images that acted as data augmentation. With this approach we reach a validation accuracy around 82%. For re-identification purpose, we added an additional classifier to discriminate the person id of the pedestrian inside the image. This helps the network to learn a way to distinguish better the different people and predict similar attributes. After this training we extract the features from the backbone for each image and we compute the distance between these context features. The closest vectors are considered to be from the same person.

2. Proposed Solutions

We started with a ResNet-50, pretrained on the ImageNet dataset. From the ResNet-50 we kept just the convolutional part and used it as a feature extractor. We opted for ResNet-50 because it achieved state of the art performances in 2020 and it is easily available in the PyTorch environment. In order to make the network resistant to noise and augment the dataset, we performed a set of image transformations on the training set. The transformations are performed randomly when returning the image from the PyTorch Dataset object. The transformations are performed on three images out of four and one time out of four the image is passed without any modification. The following transformations have been used in sequence for training:

1. horizontal flip with a probability of 50%
2. random perspective with a distortion scale of 0.1 and a probability of 20%
3. random rotation that goes from -25° to +25°
4. random erasing with a scale that spans from 0.1 to 0.3 and with a probability of 50%

To all images, we applied a normalization with mean vector equals to [0.485, 0.456, 0.406] and standard deviation equals to [0.229, 0.224, 0.225].
To get higher validation accuracy and create a network resistant to overfitting we used batch normalization and weight decay.
The network is composed with one backbone that is ResNet-50 and a set of classifiers that take as input the flattened output of the last adaptive pool layer and give an output based on the attribute that they are classifying.
For person re-identification we added a classifier to predict the person id of the image; this class is used just for training purposes.
We computed the loss with cross entropy both for attributes

and identity and combined them taking inspiration from [2]. We have optimized the parameters of the network with stochastic gradient descent algorithm with Nesterov momentum and weight decay. The learning rate is adaptively reduced over time every 5 epochs.
The annotated dataset has been splitted in training and validation set that are respectively the 80% and 20% of the total Market-1501 annotated dataset. We ensured that images of the people inside the training set are not present in the validation set.

2.1. Attributes Classification

As briefly introduced in the previous section, the attribute classification task has been approached as a multi-task learning where one network learns common features that are input for the attributes' classifiers. Each classifier is just a multilayer perceptron with batch normalization, dropout and leaky ReLU activation function that has an input size equal to the last adaptive pool layer flattened and an output size equal to the number of classes that has to predict. To enhance the classification accuracy some attributes have been grouped together and treated instead of been in a one hot encoding. Indeed, there are nine attributes that inform the color of the upper clothing, since the upper clothing can have just one of these values at a time all the *upcolor* are grouped together, same for the *downcolor*.

Regarding the fact that in the test set there are also some junk images, a possible solution to avoid to do predictions also for them and instead discard these images, could be to train a classifier in order to discriminate this kind of junk images where there is no a pedestrian inside.

2.2. Person Re-Identification

As anticipated before, we perform the person re-identification task recycling the learned ResNet-50 designed for the classification phase. Before start computing the score between query and search images, we compute a distance matrix where each entry in position *i, j* describes the distance from the image *i* and the image *j*. Each entry is the cosine distance and is computed by taking the features flattened after the last adaptive pool layer concatenated with the predicted attributes of image *i* and multiplying this resulting vector by the one transposed of image *j*. We normalize this product and subtract this to 1 to obtain a distance measure.

We also tried to follow the steps of the work [4] that retained state of the art performance on person re-identification. The approach proposed use a triplet loss. The triplet loss penalizes the network when the difference between the distance of an anchor image and a positive image and the distance between the anchor image and a negative image is less than a predefined margin. Unfortunately, with this method we were not able to produce any improvement.

3. Results

3.1. Attributes Classification

Thanks to the grouping approach of the up color and down color attributes we obtained a good boost in the accuracy. We got an accuracy with a value of 81.9% on the validation set and a 95.3% on the training set after ten epochs. We tried to continue the training for other epochs, but the validation accuracy never goes significantly above 81.9%, instead the training accuracy continue to rise, highlighting an overfitting on the dataset despite the regularization techniques that we used. This is probably due to the loss for the id classifier introduced for the re-identification task. Fig. 1 and Fig. 2 show the training and validation accuracy over the epochs.

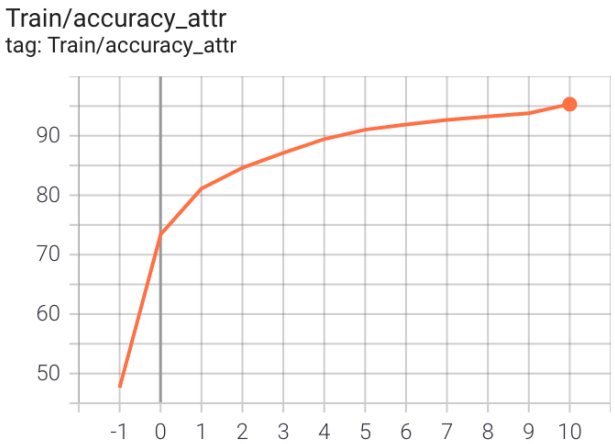

Figure 1: Curve representing the training accuracy during the ten epochs.

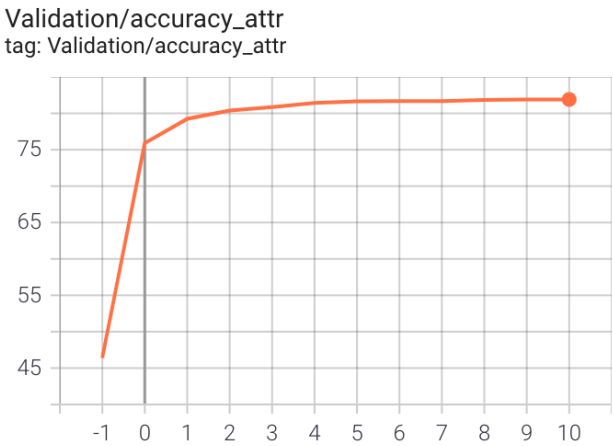

Figure 2: Curve representing the validation accuracy during the ten epochs.

3.2. Person Re-Identification

The network is trained to extract the correct features for the classification phase but also learns to distinguish different people by using an additional loss based on person classification. Computing distance on these features we achieve a mAP of 0.65. We were able to get a higher mAP on the validation set using a re-ranking technique but unfortunately it was computationally too expensive to be actually used in a bigger dataset such as the test set, so we dropped the use of this technique. It provided a boost of about 0.08 in the mAP score. The images in the retrieved set are limited to the first 30 images with a distance less than 0.5 from the query image. This is done empirically trying to have a higher mAP and reduce the number of errors.

4. Conclusions

In this work, we proposed an attribute and person recognition network able to detect with an 81.9% accuracy the attributes of a person in an image on the validation set and also able to learn a re-identification embedding. The experimental results on Market-1501 dataset demonstrate that by learning a more discriminative representation, the network achieves good re-identification performance on unseen pedestrian identities.

References

[1] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhi- lan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 2019.

[2] 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)*, pages 1116–1124, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for Image Recognition, 2015

[4] Mikolaj Wieczorek, Barbara Rychalska, Jacek Dabrowski. On the Unreasonable Effectiveness of Centroids in Image Retrieval, 2021.