COL780 Re-ID project

Rayyan Shahid Prakhar Jagwani

Abstract

Person re-identification (Re-ID) aims to identify the same Person from a variety of non-overlapping viewpoints from multiple cameras. In this project, we present a distance metric-based deep learning framework capable of Person Re-ID. A deep cosine metric learning model is kept as the baseline of our framework which uses a CNN architecture with residual blocks (BigNet) to generate feature map and uses the cosine softmax classifier during training. The baseline without augmented dataset gives a maximum mAP score of 0.861 on the provided validation set. Training the model on augmented dataset improved the maximum mAP score to 0.878. The baseline model was improved by considering a horizontal slicing of images (SliceNet1) before passing it to BigNet and the obtained features were concatenated before passing it to the cosine softmax classifier. A variation of the above model (SliceNet2) considered the slicing of feature map generated by a modified BigNet and concatenating them before passing it to cosine softmax classifier. A maximum mAP score of 0.951 and 0.953 was observed for SliceNet1 and SliceNet2 respectively.

1. Introduction

Person re-identification (Re-ID) is a well-known problem in computer vision-based surveillance. This project explores distance metric-based deep learning frameworks for Re-ID. The baseline model for our framework is based on deep cosine metric learning model [4]. The proposed model is a CNN architecture with residual layers in between. It is 15 layered network (including 2 convolutional layers in each residual block) and is therefore a relatively shallow network. This is especially advantageous for us because be have a small dataset and having a shallow neural net eliminates the requirement of pretrained deep neural networks. Moreover, training and testing times are also quick which allows a host of tweaks to the baseline to improve its performance. We call this network BigNet.

After generating the feature map, the model is passed on to a cosine softmax classifier for classification during the training process. The cross-entropy loss is minimized here which results in examples being pushed away from the decision boundary towards their parametrized mean. Thus, cosine softmax classifier becomes a sensible choice for metric-based deep learning frameworks.

We explore different improvements to the baseline model. SliceNet1 considers a horizontal slicing of the image before being passed on the BigNet. The features obtained by this network are concatenated and subsequently passed on to the cosine softmax classifier. Qualitatively, slicing of image and concatenating the features can be thought of as concatenating the local features of an image. Local features play an important role in determining the identity of a person and hence it is expected to show improvements. Another variation SliceNet2 considers slicing of the feature map generated by the last residual layer of BigNet. The sliced features are passed through the next dense layer of BigNet and subsequently concatenated before passing it to the l_2 normalization layer and then to the cosine softmax classifier. This approach is similar to AlignedReId [2], with the major difference being that we don't need a separate measure of local distance for the local features (computed by realigning the similar local features). This is because our dataset contains well cropped images of people and as a result the corresponding features of a person are highly likely to be at the same horizontal level in different images.

2. Related Work

There are a variety of metric-based deep learning frameworks which are capable of Person Re-ID. The survey paper [] discusses many of the relevant methodologies pertaining to deep learning based techniques for person Re-ID. AligendReID[2], Cosine metric based Re-ID[4] and transformer-based methods[3][1] are some of the metric-based frameworks achieving decent results. AlignedReID used both local and global features while training the model. A local distance is computed by realigning the similar local features between images and it is combined with a global distance to obtain total distance and this to which triplet loss is applied during learning. It also proposes a mutual learning approach to further improve the model. Cosine based metric Re-ID uses a shallow encoder network with

Name	Patch Size/Stride	Output size
Conv 1	$3 \times 3/1$	$32 \times 128 \times 48$
Conv 2	$3 \times 3/1$	$32 \times 128 \times 48$
Max Pool 3	$3 \times 3/2$	$32 \times 64 \times 24$
Residual 4	$3 \times 3/1$	$32 \times 64 \times 24$
Residual 5	$3 \times 3/1$	$32 \times 64 \times 24$
Residual 6	$3 \times 3/2$	$64 \times 32 \times 12$
Residual 7	$3 \times 3/1$	$64 \times 32 \times 12$
Residual 8	$3 \times 3/2$	$128 \times 16 \times 6$
Residual 9	$3 \times 3/1$	$128 \times 16 \times 6$
Dense 10		128
l_2 normalization		128

Table 1. BigNet (128×48 is the image dimension)

Name	Patch Size/Stride	Output size
BigNet		128
Cosine softmax classifier		C

Table 2. Baseline (here C = number of classes)

Name	Output size
Slicing	$K \times \frac{H}{K} \times W$
BigNet-mod1	$K \times \frac{128}{K}$
Concatenate	128
Cosine softmax classifier	C

Table 3. SliceNet1 (here K = number of slices and C = number of classes, BigNet-mod1 has the Dense 10 layer modified to return an feature vector of size $\frac{128}{K}$)

Name	Output size
BigNet-mod2	$128 \times 16 \times 6$
Slicing	$K \times 128 \times \frac{16}{K} \times 6$
Dense 10	$K \times \frac{128}{K}$
Concatenate	128
l_2 normalization	128
Cosine softmax classifier	C

Table 4. SliceNet2 (here K = number of slices and C = number of classes, BigNet-mod2 has the Dense 10 and the l_2 normalization layers removed)

cosine softmax classifier to achieve decent accuracy and usability in online scenarios. Transformer based methods are relatively recent and also show promising results in person Re-ID.

3. Methodology

The methodology followed and the detailed model and training description is provided in the following sections.

3.1. Baseline

The architecture for BigNet is shown in the table. The baseline consists of BigNet followed by cosine softmax classifier for classification during training. l_2 normalization has been applied in the final layer of encoder network (BigNet) to ensure that representation is unit length. The weights are also normalized to unit length such that $\tilde{\omega_k} = \omega_k / ||\omega_k||_2$. Then, cosine softmax classifier is stated as

$$p(y = k|r) = \frac{\exp(\kappa \cdot \tilde{\omega_k}^T r)}{\sum_{n=1}^{C} \exp(\kappa \cdot \tilde{\omega_k}^T n)}$$

3.2. Improvements

The various improvements considered are described below

3.2.1 Training

We have augmented the training dataset by adding horizontally flipped images to it. In addition to this, the images are normalized to get rid of any lighting changes that could lead to bias. This is especially important because the images are from taken from multiple cameras.

3.2.2 SliceNet1

As describes in previous sections, firstly the image is slices into K slices. Then each of this slices is passed to the BigNet to generate K number of 128 length feature vectors. These feature vectors are then concatenated and subsequently passed to cosine softmax classifier for classification during training.

3.2.3 SliceNet2

As described in the previous sections, feature map generated by the final residual layer of BigNet (Residual 9) is sliced into K segments. These slices are then passed through the Dense 10 layer and then concatenated. It then passes through l_2 normalization layer and then finally passed to the cosine softmax classifier for classification during training.

4. Results

Model	Rank-1	Rank-5	MAP
Baseline	0.929	1.000	0.861
Baseline$_{AUG}$	0.857	1.000	0.878
SliceNet1	0.964	1.000	0.951
SliceNet2	0.964	1.000	0.953

Table 5. Performance comparison

All the models have an almost perfect Rank-5 score. This is because many of the images in the query set, were present in the gallery. Since, the number of query images were low, Rank-1 varied greatly between training runs. For the given validation set, `MAP` seems to be the most reliable metric.

5. Analysis

The improved model seems to work slightly better in cases where the persons look mostly similar, except for some local feature, which couldn't get enough representation in the feature vectors.

Here, we compare SliceNet2 and Baseline$_{AUG}$ on some query images.

5.1. Improvements from Baseline

In Figure 1, the base model fails to differentiate between the persons `12` and `32`, which look very similar from the back. SliceNet, which depends more on local features, fares better here, probably because of the easily differentiable head and torso.

In Figure 2, the persons `73` and `88` are wearing similar shirts. The base model ranks a picture of `88` the highest, probably due to this reason. The improved model, based on local features, doesn't miss the difference in the hands.

5.2. No Improvements from Baseline

In Figure 3, both the models perform poorly, because the person's bag has been occluded, and `12` looks similar.

In Figure 4, the improved model performs poorly. Local features don't work well here, probably because the top quarter of the images look similar.

6. Conclusion

The baseline model gives decent results in many scenarios and achieves a mAP score of 0.861. This score is improved by using an augmented dataset during training and mAP score of 0.878 is achieved. Using the model SliceNet1 and SliceNet2 further improves the performance as local features of a person are also kept track of. SliceNet1 and SliceNet2 achieve a mAP score of 0.951 and 0.953 respectively. There are still certain scenarios where the improved models fail as well. These include cases where certain distinctive features of a person are occluded in some of the frames.

7. Weights

Weights can be found here.

8. Contributions

1. Generating custom dataset and dataloader - Rayyan
2. Training - Prakhar
3. Baseline model - Rayyan
4. Improvements - Prakhar
5. Report - Prakhar and Rayyan

References

[1] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification, 2021. 1

[2] Hao Luo, Wei Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, and Chi Zhang. Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognition*, 94:53–61, 2019. 1

[3] Charu Sharma, Siddhant R. Kapil, and David Chapman. Person re-identification with a locally aware transformer, 2021. 1

[4] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 748–756. IEEE, 2018. 1

Figure 1. SliceNet works better than Baseline here.

Figure 2. People with similar shirts.

Figure 3. Both models are confused, because the bag is occluded

Figure 4. Local features are at a disadvantage here.