CS230

Tackling Occlusion in Person Re-identification

Soutik Chakraborty
Stanford University
soutikc@stanford.edu

Boxiao Pan
Stanford University
bxpan@stanford.edu

Pragya Mishra
Stanford University
pragmi@stanford.edu

Abstract

In this paper, we improve the state-of-the-art Person Re-ID model on the edge cases of occluded person images, by incorporating a multi-task loss and training jointly on both occluded and non-occluded images. Since there is no publicly available occluded Person Re-ID dataset, we implement an occlusion simulator to automatically generate occluded images. Experimental results show that our model almost doubles the CMC10 score on occluded images compared to the original model, while doesn't have a big drop on non-occluded images. We also carry out qualitative analysis via saliency maps, which shows that our model successfully learns to avoid paying attention to the occluded areas.

1 Introduction

Person Re-identification (Person Re-ID) is a challenging task to retrieve a given person among all the gallery pedestrian images captured across different security cameras. The main challenges for person Re-ID come from large variations on persons such as pose, clothes, background clutter, etc [1]. The task can be even harder when we take occlusion into consideration, which means some part of the person to be identified might be occluded (by objects, other people, etc), which is often the case in reality but has not been paid much attention by researchers. This problem is defined as the occluded Person Re-ID. In this project, we improved on the state-of-the-art Person Re-ID model (i.e., Aligned Re-ID [2]) by incorporating a multi-task loss [3], and trained the new model jointly with non-occluded images and occluded images, which greatly enhanced its performance on occluded person images without losing too much accuracy on non-occluded images.

2 Related work

Currently, Person Re-ID is often solved in two separate steps. The first step is to extract robust features that are invariant to the changes of illumination, clothes, viewpoint, etc. While the second step is to design a metric so that the distance between images of the same identity is small while big for images of different identities.

The Person Re-ID community has made great progress in both of these areas. In terms of feature extraction, [4] presents a multi-scale network architecture called MSCAN to capture fine-grained as well as context-aware information. Besides, it also proposes another network called STN to adaptively localize body parts, then send into MSCAN together with the whole body feature to obtain final results. [5] proposes a Spindle Net to leverage human body structure information by using a body regional proposal network. They then use a multi-stage pooling framework followed by a tree structured merging network to select the most useful features.

While for metric learning, triplet loss is by far the most popular loss term in Person Re-ID. Our baseline model [2] designs a global distance together with a local distance and applies triplet loss

CS230: Deep Learning, Fall 2018, Stanford University, CA.

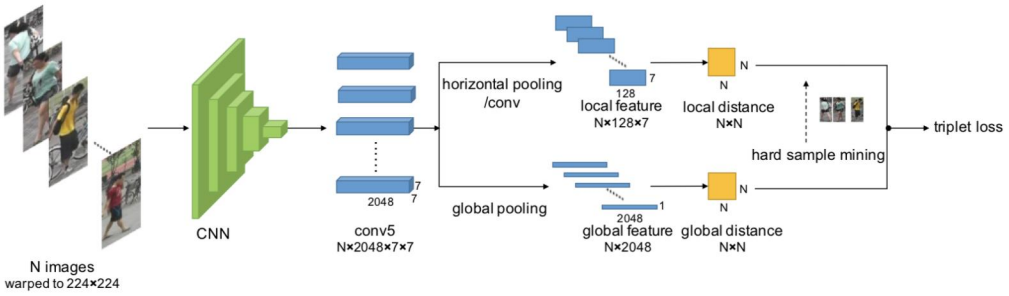

Figure 1: Aligned Re-ID Model Architecture [2]

based on these two distances (a detailed explanation will be provided in the following section). Other works include [6] which proposes a patched-based metric learning approach that greatly decreases the dimensionality of the extracted feature vectors, and [7] which introduces a quadruplet loss function to further decrease intra-class variation while also increase inter-class distance.

Although we have seen many works in this exciting field, there has not been much attention paid to the Occluded Person Re-ID problem. [3] deals specifically with the problem of Occluded Person Re-ID and is also the main work we refer to. They propose a multi-task loss to take the source distribution (occluded or non-occluded) of query image into consideration, thus enabling the model to learn occlusion-robust features. But they only tested their method on the framework they designed specifically for this task without incorporating into existing networks, thus it is unclear about the generalization ability of their approach. Moreover, they only tested their approach on occluded Person Re-ID datasets (none of which is released), so we don't know if this method would cause serious drop on non-occluded images. We solve these problems in our work, and we also further explore the effect of different occlusion type, which may provide novel insight into how Person Re-ID model deals with occlusion.

3 Baseline Model

In this section we provide a detailed description of the baseline model we use, which is the current state-of-the-art Person Re-ID model named Aligned Re-ID [2].

3.1 Model Architecture

The CNN shown in Figure 1 is a ResNet50 model pre-trained on ImageNet, which is used to extract features from input images. It is followed by a conv layer of size $2048 \times 7 \times 7$. This layer feeds into two subsequent networks to generate local features and global features which are then used to calculate the Triplet loss function.

3.2 Loss Function

3.2.1 Global Distance

Global distance is calculated as the L_2 distance of the global features between two images, in which the global feature is obtained by directly applying global pooling on the feature map.

3.2.2 Local Distance

First, horizontal pooling is applied, which is essentially global pooling in horizontal direction on the feature map obtained from the convolution layer. This reduces a $C \times H \times W$ feature map to $C \times H$. Post that a 1×1 convolution is applied to further reduce the channel number from C to c. This way each local feature represents a horizontal part of the image.

Given the local features of two images, $F = f_1, ..., f_n$ and $G = g_1, ..., g_n$, we calculate a distance matrix D where each element

$$d_{i,j} = \frac{e^{(||f_i - g_j||_2)} - 1}{e^{(||f_i - g_j||_2)} + 1} \tag{1}$$

The local distance is then defined as the total distance of the shortest path from $(1, 1)$ to (H, H) in D.

3.2.3 Triplet Loss

Triplet loss with hard example mining together called TriHard loss is defined in [8]. For each sample, based on the global distance, the most dissimilar and most similar one are chosen to obtain a triplet. After calculating the global and local distances, triplet loss is applied to obtain the final loss function.[2]

4 Methods

In order to help our model learn occlusion-robust features, we first add an additional occlusion label to each image, which denotes whether it is occluded or not (0 for occluded, 1 otherwise). Then we introduce an OBC loss (which will be explained later) and incorporate it into the original triplet loss to form what we call the multi-task loss. Finally, we train this new model on both occluded and non-occluded images.

4.1 Multi-task Loss

First, we define the OBC loss as follows:

$$L^o(\hat{y_i}', y_i') = \sum_{c=1}^{C} (y_i' = c) \log(e^{\frac{\hat{y_i}'}{\sum_{c=1}^{C} e^{\hat{y_c}'}}}) \tag{2}$$

Where y_i' is the occlusion label we defined earlier. We add another fully-connected layer to predict this label.

Then combine it with the original triplet loss, we obtain the final multi-task loss as:

$$L = \alpha L^T + (1 - \alpha) L^O \tag{3}$$

Where L^T denotes the triplet loss and α is a hyper-parameter which balances the two losses.

4.2 Occlusion Simulator

The algorithm below was used to create the occluded images out of the original Market1501 dataset. It should be used only on query images while not on gallery images.

Algorithm 1 Occlusion Simulator

1: **function** OCCLUSION SIMULATOR($image, patch\_size$)
2: $average\_color \leftarrow mean$(All pixel values in the image)
3: $occlusion\_mask \leftarrow numpy.ones(50, 50, 3) * average\_color$
4: $occluded\_image \leftarrow apply(image, occlusion\_mask)$
5: **return** $occluded\_image$

Where $apply(image, occlusion\_mask)$ applies $occlusion\_mask$ on $image$ at a random location.

5 Dataset

5.1 Market1501

The Market1501 dataset is a standard dataset used for person re-identification. Market1501 dataset is a collection of 32,688 annotated images collected from the security cameras outside Tsinghua University [9].

5.2 Occluded Market1501

We applied our occlusion simulator on the original Market1501 dataset to get the occluded version of it. We give some examples of the non-occluded and occluded images in Figure 2.

6 Experiments and Discussion

6.1 Metric

We use CMC10 (Cumulative Match Characteristic over 10 images) as a metric to determine the effectiveness of the model. CMC10 is the percentage of images in the top 10 images ranked by the model for a query image which are actually true matches for the query image.

6.2 Quantitative Results

We compared our model's performance against the original Aligned Re-ID's performance by testing on both Market1501 and Occluded Market1501 datasets. The results are listed in Table 1.

Model	**Market1501**	**Occluded Market1501**
Aligned Re-ID	**95.61%**	18.20%
Aligned Re-ID + OBC loss	77.40%	**30.94%**

Table 1: Quantitative experimental results.

From the results we can see that the CMC10 score of our model on occluded images almost doubles that of the original model, which suggests that our model is more robust to occlusion. Moreover, the performance on non-occluded also drops, which is reasonable since occluded images act as noise during training. And compared with the improvement on occluded images, we deem this non-significant drop fairly acceptable. In practice, we can tune the hyper-parameter α to balance the accuracy on non-occluded and occluded images.

6.3 Qualitative Results

Saliency maps represent the probability of visual attention captured inside a neural network. They are being used to understand what is being learned in various layers of network as well as indicating which areas of the image are getting the most attention by the network. We use the Grad-CAM [10] to visualize the activation map from the last conv layer. We show the results in Figure 2. As expected, Aligned Re-ID pays a lot of attention to the occluded region resulting in poor performance. On the other hand, adding OBC loss lets our model pay less attention to occluded areas, which is exactly what we expected.

6.4 Analysis of Different Occlusion Type

We experimented with different types of occlusions during our experimentation. We started off with a completely black patch of occlusion. But we found that with this kind of occlusion the model pays all its attention to the occluded area (due to space limit we can't insert images to illustrate). We then moved to a more natural representation of occlusion by setting the occlusion value to be the average pixel value of the original image. The difference between different occlusion type comes very interesting and we will investigate the deeper reason of it in future work.

6.5 Analysis of α

In order to balance the OBC loss with the original triplet loss (which acts as the identity loss), we used a hyper-parameter α to do the trade-off. After performing a grid search over different α values, we got the best validation set performance by setting α to 0.8, which means that the OBC loss should pose auxiliary influence while the identity loss (or triplet loss in our case) takes the main role.

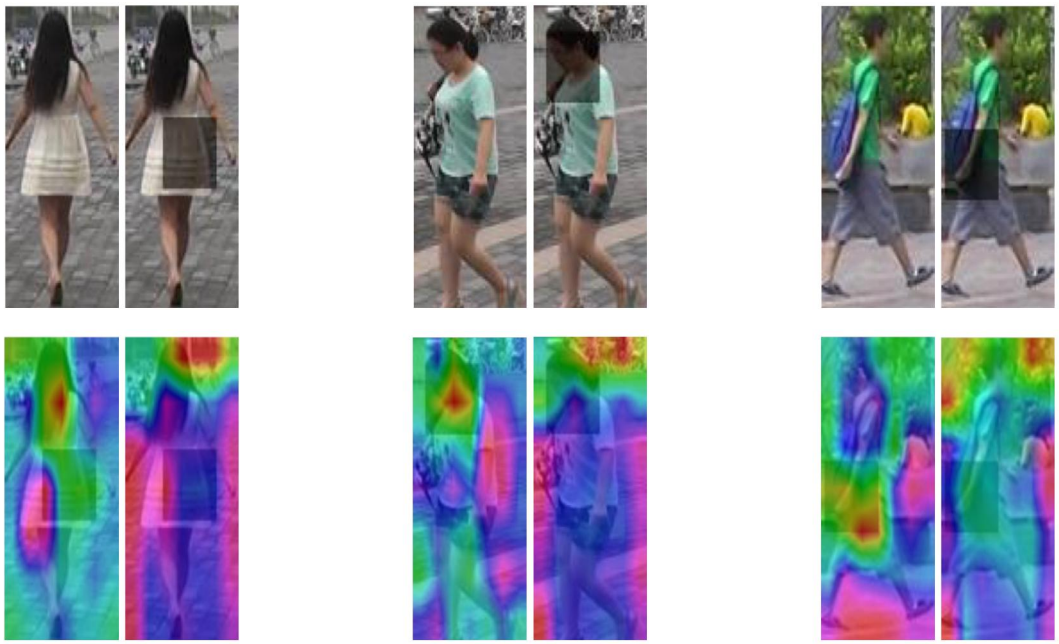

Figure 2: Visualization using images from three different identities. Each 4-image cluster belongs to 1 identity, in which the top left corner is the original image, while the top right corner is its randomly occluded version. The bottom left corner shows the saliency map using the original Aligned Re-ID model , while the bottom right corner denotes the saliency map using our new model.

7 Conclusion and Future Work

In conclusion, after incorporating OBC loss, our model performs much better on occluded images than the original Aligned Re-ID model without dropping too much accuracy on non-occluded images. Also from the saliency maps, we can see that our model has learnt to put less attention on the occluded areas, which is exactly what we expect. The occlusions are applied randomly and are of the average value of all pixels, which confuses the model so much as to where to pay attention to and where not to. Given this, the results are indeed very promising. Moreover, our work proves the generalization ability of this approach in dealing with occlusions, thus can be further researched and incorporated into other existing Person Re-ID frameworks.

In the future, we will carry out the following ideas:

1. Adopt a GAN instead of a single fully-connected layer to predict the occlusion label, which should be better at helping network learn features that are more robust to occlusion.
2. Gather occluded images in a realistic setting rather than automatically generating occlusions.

8 Contributions

In this section we describe the work done by each team member for this project. Every member spent considerable time reading literature and understanding the problem at hand:

Soutik Chakraborty Soutik worked on occlusion simulator, adding occlusion labels, tunning hyper-parameters and running experiments. He also helped in debugging the code base for errors.

Boxiao Pan Boxiao took charge of research ideas throughout the project, as well as modifying the model. Boxiao also provided valuable insights into solving performance bottlenecks.

Pragya Mishra Pragya took the part of generating saliency maps. She also helped in running experiments and setting up the experiment environment.

References

[1] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," 2018.

[2] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[3] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," *arXiv preprint arXiv:1804.02792*, 2018.

[4] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 384–393, 2017.

[5] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 907–915, 2017.

[6] S. Bak and P. Carr, "Person re-identification using deformable patch metric learning," *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, 2016.

[7] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2017.

[8] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017.

[9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization.," in *ICCV*, pp. 618–626, 2017.