在有限計算單元上提升行人重識別的推論速度之研究

A study of Improving person re-id inference speed on limited computation resources

組別：A99　組員：王昊文 、陳博暐　指導教授：孫民

Abstract

Person Re-Identification has been a widely studied topic in the computer vision field, following the prosperity of the Convolutional Neural Network. Person ReID can be extremely useful for public safety and surveillance cameras, as it increases the performance of matching identities across different cameras.

Accuracy is important in the field of person ReID. However, in general usage, we don't have powerful computers installed on surveillance cameras, hence the importance of calculation consumption must not be neglected. Although the accuracy for most of the benchmarks have plateaued(most of them have accuracy over 95% rank-1), there are still hundreds of new models being proposed every year, most of the state-of-the art Person ReID models seem to have complicated architectures and training with multi-branch features in order to squeeze out a little more accuracy and beat the benchmarks, the down-side of using complex model design is making its inference performance terrible when implementing on general surveillance camera systems[1]. In practical usages, over 90% rank-1 is already reasonable.

In this project we are working with the CarePlus healthcare system, which aims to monitor the health status of the elderly who are alone at home. The system mainly uses embedded systems, with a camera. ReID is essential in the CarePlus system, we use ReID to monitor in-door activities of the elderlies. However most Re-ID models are too heavy for the hardware system, hence we seek to find a solution to improve the inference speed of the Re-ID module of the Carplus healthcare system. We will try training a light-weight teacher student model, also try more simple networks, and also try different sorting algorithms and distance functions. We will combine these methods and compare the effectiveness of each method. With various combinations of these methods we successfully reduced the inference time on the careplus system as compared to the strong baseline model OSNet.

Keywords: Person-ReID, surveillance cameras, model architecture, accuracy, inference performance, lightweight hardware systems, teacher-student model, sorting algorithms, distance function.

摘要

隨著卷積神經網絡(Convolutional Neural Network)的興起，行人重識別(Person ReID)成為了計算機視覺領域的一個廣泛研究的主題。 Person ReID 對於公共安全和監控攝影機的實際應用中有著不可撼動的地位，因為它可以提高跨攝影機身份匹配的性能。

準確度在 ReID 之中極其重要。但是，在一般情況下，監視攝影機上往往沒有強大的運算能力，因此，計算消耗就顯的非常重要。儘管大多數基準測試(benchmarks)的準確度已達到穩定水平（大多數基準精度超過 95%），每年仍有上百種不同的模型被提出，大多數最新提出的 Person ReID 模型具有複雜的模型架構並使用多分支功能的架構，以提高準確度並超越最先進的的基準精度(benchmark accuray)。正因為如此，在一般監視攝影機系統上應用這些 ReID 模型的時候，使用非常複雜的設計雖然讓獲得更高的準確度，但其推論速度(inference speed)變得不佳。在一般應用中，超過 90% rank-1 準確率已經是令人滿意的。

在這個專題中，我們使用 CarePlus 醫療保健系統，該系統旨在監視獨自一人在家中的老年人的健康狀況。該系統主要使用帶有攝影機的嵌入式系統。 ReID 在 CarePlus 系統中至關重要，我們使用 ReID 來監測老年人的室內活動。但是，大多數最先進的 Re-ID 模型對於硬體系統來說負擔太大了，因此在這項專題中，我們希望提高 Carplus 保健系統的 Re-ID 模組的推理速度。我們的主要目標是在準確性和推理速度之間取得平衡。我們將嘗試訓練一個輕量級的師生(teacher-student model)模型，嘗試更簡單的網絡，並嘗試不同的排序算法和距離函數。我們將結合使用這些方法，並比較每種方法的有效性。通過這些方法的各種組合，比起強大的基線模型 OSNet 相比，我們成功減少了 careplus 系統上的推理時間。

關鍵字: 行人重識別、 監控攝影機、模型架構、準確率、推理速度、輕量硬體系統、師生模型、排序演算法、距離函數。

INTRODUCTION

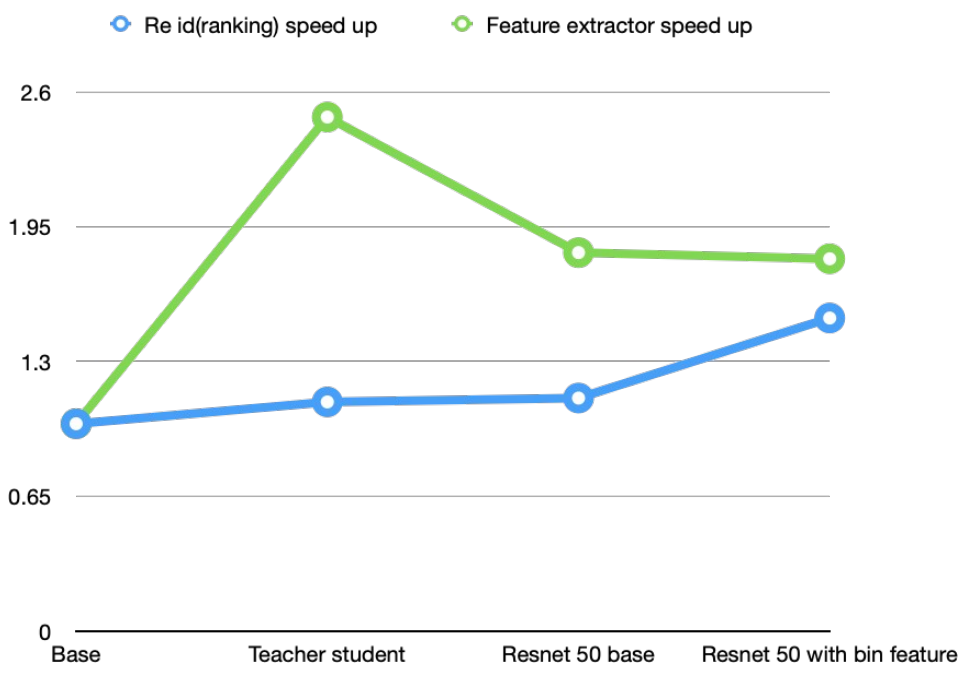
The goal of Person Re-ID is as follows: Given a query person-of-interest, we want to determine whether this person has appeared in another place at a distinct time captured by a different camera. We store a person captured by another camera inside a gallery set, usually our query person-of-interest encoded as a feature encoding via a deep model, we then compare the feature encoding with our gallery set, usually with cosine distance, or Euclidean distance. Finally, we construct a rank list, ranking the person inside the gallery set, from the most similar to the least similar.

The field of Person Re-ID has evolved rapidly these years, along with the fast-paced computer vision community, various high-quality person Re-ID based dataset was published

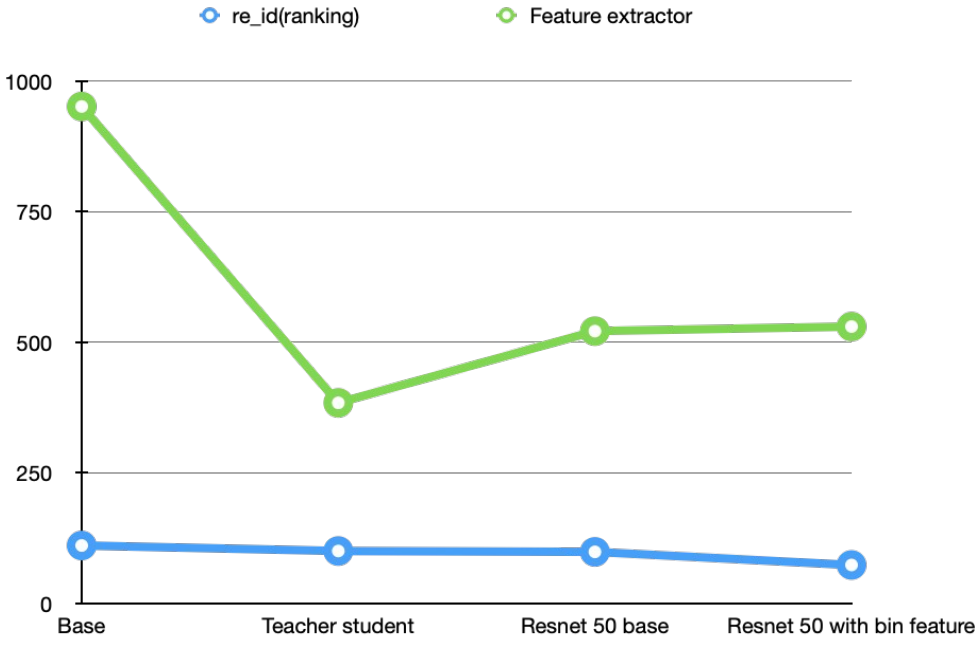
for researchers, famous datasets like Market-1501, DukeMTMC-reID, CHUK03, most of these have over 90% rank-1 benchmark performance with the state-of-the-art models. Since 2014 to now, although the accuracy for most benchmarks have plateaued, there are still countless models being proposed each year, most of them use extremely complex design and unique training tricks, making it hard to gain the intuition and validate the correctness mathematically of the model. On the other hand, it is hard to validate the results of the paper if the author decides not to make their codes open-sourced. Not only the correctness, these complex designs make it nearly impossible for practical usage, for example, a small embedded machine with limited computation resources and a light-weight camera.

In this project, we are collaborating with the Careplus health-care project, which aims to monitor the health status and the activity of an elderly at home. Person Re-ID is extremely important in this project, it monitors the activity of the elderly, we can get a descriptive sense of the elderly's habit, activity area, and further increase the safety equipment of that area. The monitor system runs on a NVIDIA Jetson Nano module, with a fish-eye camera.

In order to increase the inference speed, We first tried a lightweight model trained by the teacher-student method plus ambiguous distribution tears, but performance wasn't too good. we then use the suggestion given by [1], by using a clean fresh model with several simple tweaks, we improve the inference speed massively compared to the previous SOTA model OSNet used in the Careplus project without losing too much accuracy. To improve even more, we include the method suggested by [2] and combine the open-source Re-ID SOTA models framework maintained by [3], we encode feature into binary codes to further improve the re-ranking process, traditionally cosine and Euclidean distance is used, we now uses hamming distance and counting sort to further improve the rank list sorting process. We successfully reached 93% of rank-1 accuracy and immensely improved the inference process. In order to reduce the computation complexity, we have tried to change model architecture, use different training methods and different loss functions. By changing the model architecture, We didn't achieve a good performance model with light weight computation. We tried the data distillation method to learn the data distribution of the big model. We use 2 stacked OSBlocks (original OSNet stack 6 OSBlocks) as student models. We learn the feature distribution with temperature t = 1.8, it will reduce 79% computation of the teacher model. It achieved 48.6% mAP and 71.1% rank-1 accuracy compared with the original OSNet 77.4% mAP and 91.6% rank-1. Then we think that maybe the model is learning an ambiguous data distribution, we just need to tear that ambiguous distribution part in order not to drop accuracy. We use triplet loss to tear the ambiguous data distribution which gives us 62.2% (+13.6%) mAP and 81.7% (+10.6%) rank-1.



Speed up



Each time component

Conclusion

In this project, we seek to look for a solution to boost the performance when implementing ReID models on hardware systems. Although there are lots of new SOTA models being proposed each year, the design of the models are complex. We purpose that if we want to achieve fast inference speed, *the symmetry and the simplicity of the model is the key to really improve the performance on hardware systems*, this generally makes optimizing and parallelization better, we have shown this proposal using the baseline OSNet and ResNet-50 based models.

Although the teacher-student model on OSNet didn't perform well at accuracy, we suppose that it is because we took out too much of the original network, we *must not neglect the performance increase during feature extraction*. For further work, we may try to make a ResNet-34 model and use ResNet-50 as our base model to train a teacher-student model. If we can prevent the accuracy from dropping much, it might greatly improve the inference speed. It is important that our student model must also have a symmetric, layer repeated design for better hardware computation.

As for ReID specific, the ranking procedure can also take up a lot of time. We propose that in ReID lightweight hardware systems we should learn binary encoded features, as that greatly improves the ranking process. On 117 dataset and current Careplus ReID module design, there are not many ranking calls, however for practical usages, binary encoded features can greatly reduce the computation requirement on small, lightweight computers. For using these design principles, we reduce the total ReID module inference time on the Careplus system for 36.5% as compared to the baseline OSNet model.

心得

王昊文：

剛開始加入有關 computer vision 相關的實驗室，單純只是因為自己好像對電機所有領域都不太有吸引力，而這個領域沒接觸過想要勇敢嘗試，後來隨著大三修了很多相關領域的課程，覺得這個領域真的太有趣了，有很多接觸最先進 AI 相關的技術的機會，也確立了自己以後我相關領域發展的目標。因為我是電機系的起頭的時候稍微吃力了一些，本身之前也沒有太多接觸大型軟體專案的經驗，幸好我也很罩的組員陳博暐還有超強的實驗室學長的指導與建議，讓我很快的進入狀況。感謝實驗室還有教授願意帶領我們完成這次的專題！

陳博暐：

這一年的專題讓我對 machine learning and computer vision 這領域的相關知識更加了解。在一開始，因為背景知識的不足，在專題的一開始像是無頭蒼蠅一般的胡亂嘗試，在閱讀更多的論文與跟學長教授更細緻的討論後，對該領域的知識更加

熟悉也得以更加順利的進行專題研究。在此要感謝實驗室學長與教授這一年以來非常有耐心的給我們方向和給我們不少的建議，還有我的組員王昊文，若沒有他的協助與討論，我們沒辦法完成這個專題。

Reference

[1] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[2] Guan'an Wang, Shaogang Gong, Jian Cheng, and Zengguang Hou. Faster Person Re-Identification. In *European Conference on Computer Vision(ECCV)*, 2020.

[3] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng and Tao Mei, JD AI Research. FastReID: A Pytorch Toolbox for General Instance Re-Identification.

[4] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *ICCV*, 2019, pp. 3702–3712.

[5] Mang Ye, Jinbing Shen, Gaojie Lin, Tao Xiang Ling Shao and Steven C. H. Hoi, "Deep Learning for Person Re-identification: A Survey and Outlook" *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, 2020

[6] Chen, J., Wang, Y., Wu, R.: Person re-identification by distance metric learning to discrete hashing. In: IEEE *International Conference on Image Processing*(ICIP). pp. 789–793 (2016)

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

[8] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[9] Geoffrey Hinton, Oriol Vinyals, Jeff Dean. Distilling the Knowledge in a Neural Network. NIPS 2014 Deep Learning Workshop