# Person Re-Identification using LA Transformer

Somanshu Singla
2018EE10314
ee1180314@iitd.ac.in

## Abstract

*Person Re-Identification is a useful problem in the computer vision domain. In this work I have experimented with the Locally Aware transformer model[8] and have proposed some improvements. I have suggested improvements by use of different loss functions, change in model architecture and modification in image retrieval pipeline. Using the improvements the validation set score obtained is: Rank1:* ***1.000****, Rank5: 1.000, mAP:* ***0.975****. The code will be released at: Singla17/PersonReID.*

## 1. Introduction

Person re-identification is a well known problem in Computer Vision. In this problem we try to associate images of same person taken from different cameras. This problem finds it's use in intelligent video surveillance.

### 1.1. Locally Aware Transformer

The model I chose as the baseline is the Locally Aware Transformer called as LA Transformer. LA-Transformer combines vision transformers[5] with an ensemble of FC classifiers that take advantage of the 2D spatial locality of the globally enhanced local tokens [8].

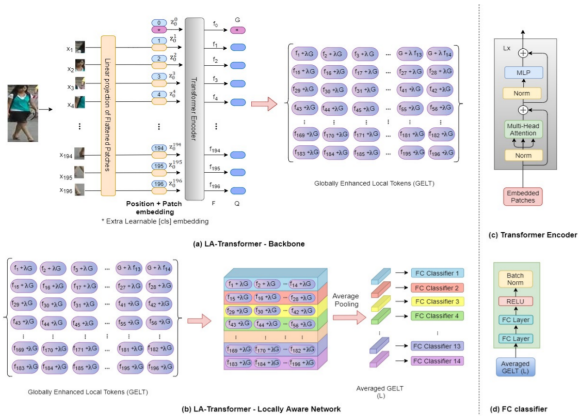


Figure 1. Architecture of LA Transformer[8]

### 1.2. Comparison with other baselines

I looked at all the baseline options available to us AlignReId[12], TransReId[1] and Deep Cosine Metric Learning[10]. AlignReId and Deep Cosine Metric based ReID use CNN as the baseline whereas TransReID uses a ViT baseline similar to LA Transformer. The reason for choosing LA Transformer was that it showed very promising results in comparison with the other methods, moreover attention based architectures such as ViT have not been explored fully in the vision domain thus making LA Transformer a good candidate to improve upon.

| Model | Market-1501 | | CUHK-03 | |
|---|---|---|---|---|
| | R1 | mAP | R1 | mAP |
| LA Transformer | **98.27** | **94.46** | **98.7** | **96.4** |
| TransReID | 95.2 | 89.5 | - | - |
| AlignReID | 91.8 | 79.3 | 92.4 | - |
| Deep Cosine Metric Based | 79.10 | 56.68 | - | - |

Also as we see in the the table above that LA Transformer has claimed the best results on both the datasets among the baseline candidates.

## 2. Prior Work Done

A lot of work has been done on this problem where people have tried a large variety of CNN based architectures and there has been some use of Transformers as well. We can get an idea of current work done from [8] which states: "For many years CNN based models have dominated image recognition tasks including person re-ID. A vast body of research has been performed to determine the best strategy to extract features using CNNs to address issues like appearance ambiguity, background perturbance, partial occlusion, body misalignment, viewpoint changes, and pose variations, etc. In [7] proposed a PoseSensitive Embedding to incorporate information associated with poses of a person in the model, In [11] used a Graph Convolution Network[4] to generate a conditional feature vector based on the local correlation between image pairs, In [3] authors used global channel-based and part-based features, In [13]

authors used global pooling to extract global features and horizontal pooling followed by 1 × 1 CNN for local features. CNN based methods have led to many advances in recent years and are continuing to be developed for person re-ID."

In [1] vision transformers were applied to the domain of Person Re-Identification for the first time and they also achieved results comparable to state of the art CNN based methods.

## 3. Improvements Made

I have made improvements to the standard LA Transformer model in three aspects:

1. *Loss Function*: Experimented with various loss functions other than Cross-Entropy

2. *Model Architecture*: Changed the way features are extracted after passing from the ViT baseline

3. *Modification to Image Retrieval Pipeline*: Changed the prediction pipeline to make it more robust to variations in query images

### 3.1. Loss Function

The original LA Transformer uses simple cross entropy loss function but there are a variety of other kind of loss functions which are useful in our context and I have used two such loss functions: AM Softmax, Triplet Loss.

#### 3.1.1 AM Softmax

AM Softmax[9] stands for additive margin softmax and is a modification on the top of standard Softmax-CrossEntropy duo used in classification tasks. The AM Softmax loss function tries to introduce a margin in the decision boundary between different classes and aims to maximize the inter class distance and minimize the intra class distance. It has been shown to be the most effective variant of softmax-crossentropy loss functions among L-Softmax, A-Softmax.

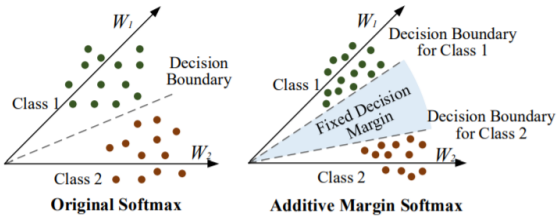


Figure 2. Comparison of Softmax and AM Softmax[9]

#### 3.1.2 Triplet Loss

Triplet Loss[2] is a very popular choice of loss function in person re-identification tasks. The basic idea of triplet loss is that the distance between two images of same classes should be less then the distance between images of different classes by a fixed margin. In essence it also tries to reduce the intra-class distance and maximize the inter class distance.

I have used the sum of AM Softmax and Triplet loss as my loss function although the final aim of both losses is similar they establish the goal by different means wherein the AM Softmax tries to improve the features by introducing a margin in decision boundaries whereas the triplet loss tries to do so by introducing margin in the distance between the data points.

### 3.2. Model Architecture

I modified the architecture of the locally aware network part of the LA Transformer:

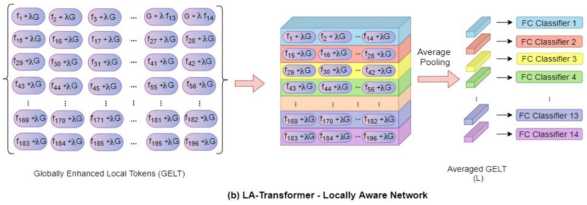


Figure 3. Locally Aware Network[8]

In this if we take a look at the pooling part we can see that they pool along the x direction because the image patches which have lead to the outputs which are in the same row are adjacent in the actual input image and they make use of that information, but what they have not used the fact that the features which are present in the same column in the feature matrix of Locally aware network come from the patches which were vertically adjacent in the input image. So to make use of that information we also pool along the y-direction and finally take the average of the features obtained along the x and y direction to get out final features.
Formally:
In the original paper:

$$Feature = AvgPool(f_1+\lambda G, f_2+\lambda G \dots f_{14}+\lambda G) \quad (1)$$

Improved:

$$\begin{aligned} Feature =& (AvgPool(f_1 + \lambda G, f_2 + \lambda G \dots f_{14} + \lambda G) \\ &+ AvgPool(f_1 + \lambda G, f_{15} + \lambda G \dots f_{183} + \lambda G))/2 \end{aligned} \quad (2)$$

This helps the feature to extract more local information as the y adjacent patches are also local and for a case where

we detect humans whose height is larger than width the information in y adjacent patches is equally important if not more.

### 3.3. Modification to Image retrieval pipeline

Whenever we are training we apply various data augmentations such as flipping transformations to make the data more generic and to make the learning more effective but no such technique is applied to retrieval pipelines. In this I also pass the flipped version of the query image to the model and the final feature which I use to retrieve image from the gallery is the average of features of the original image and the flipped image thus making the retrieval pipeline more robust.

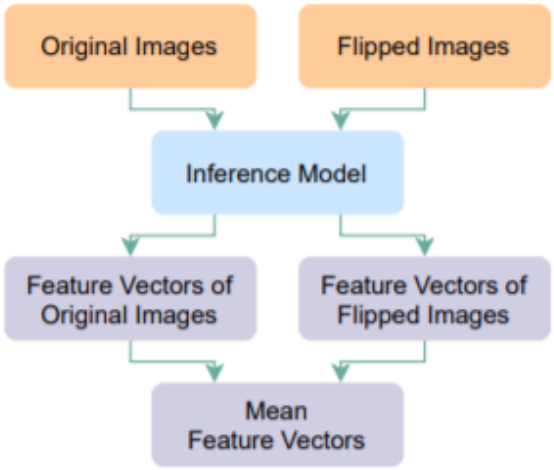


Figure 4. Retrieval Pipeline using Flipping[6]

## 4. Comparison of Baseline and Improved Model

The model weight file links:

- For improved model can be found: here.
- For baseline model can be found: here.

The first thing which we can see are the metrics on the val set for both the models:

- Baseline:
  Rank1: 0.929, Rank5: 1.000, mAP: 0.908
- Improved Model:
  Rank1: **1.000**, Rank5: 1.000, mAP: **0.975**

The scores[1] clearly indicate that the improved model has performed better on all the metrics

[1]The difference in score numbers is larger than what it should have been as the val set is extremely small and even the difference on 1/2 queries can create a huge difference.

An Example where baseline gets it wrong and improved model predicts correctly:

Figure 5. Query

Figure 6. First Retrieved Image

Figure 7. Example of retrieval pipeline for baseline model

Figure 8. Query

Figure 9. First Retrieved Image

Figure 10. Example of retrieval pipeline for improved model

As we can see in 7 the retrieved image by the model is of different person whereas the improved model is able to retrieve image of the correct person from gallery.

We can also compare the models based on the features learned by the models and the images below will help us do so:

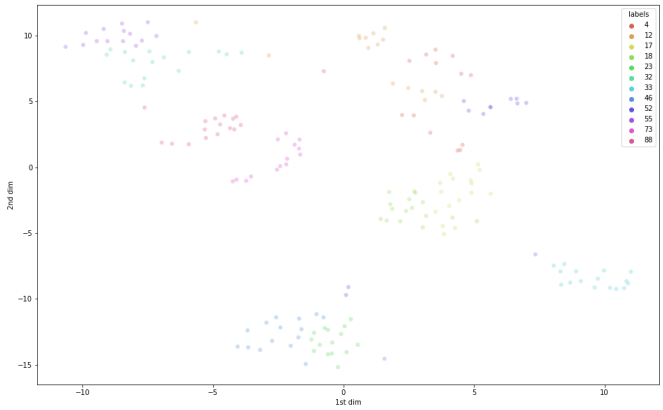


Figure 11. Feature Vector for val set gallery using baseline model

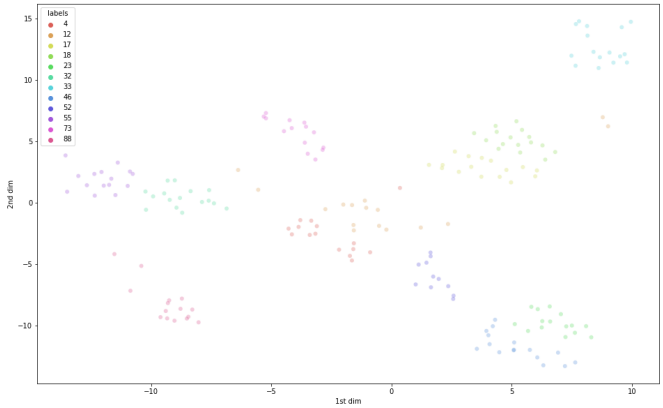


Figure 12. Feature Vector for val set gallery using improved model

Figure 13. Feature Vector plot using tSNE

In this[2] I have plotted the first two dimensions of the val set gallery images using tSNE and as we can see from the plots that for the improved models the features are more separable with one class taking lesser area and less mixing between two classes, which makes it much easier for the model to make decision boundaries and hence improving the performance.

## 5. Ablation Study

In this section I have tested the models by removing one of the improvement techniques and have tested it on the validation set to see the effect of each of the improvement technique proposed. The Result Metrics were:

[2]here the labels are the category numbers of the datapoints in the val set

- LA-Transformer with y-pooling:
  Rank1: *0.964*, Rank5: 1.000, mAP: *0.947*

- LA-Transformer with y-pooling, AMS+Triplet loss:
  Rank1: *1.000*, Rank5: 1.000, mAP: *0.971*

- LA-Transformer with y-pooling, AMS+Triplet loss, Flipping: Rank1: 1.000, Rank5: 1.000, mAP: *0.975*

As we can see from the scores that each of the proposed changes have helped in increasing the score thus showing the usefulness of these ideas.

## 6. Conclusion and Future Work

In this project I was able to implement the LA Transformer model and have proposed some ideas using which the performance of the LA Transformer model can be improved. The usefulness of the ideas has been established in the comparison with baseline and ablation study sections.
AS part of future work the baseline ViT model can be changed, the current baseline for LA Transformer is the ViT small, ViTs with greater number of layers and larger capacity can be tried out.

## References


[1] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification, 2021. 1, 2

[2] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification, 2017. 2

[3] Fabian Herzog, Xunbo Ji, Torben Teepe, Stefan Hörmann, Johannes Gilg, and Gerhard Rigoll. Lightweight multi-branch network for person re-identification, 2021. 1

[4] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017. 1

[5] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 1

[6] Xingyang Ni and Esa Rahtu. Flipreid: Closing the gap between training and inference in person re-identification, 2021. 3

[7] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, 2018. 1

[8] Charu Sharma, Siddhant R. Kapil, and David Chapman. Person re-identification with a locally aware transformer, 2021. 1, 2

[9] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 2

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

[11] Fufu Yu, Xinyang Jiang, Yifei Gong, Shizhen Zhao, Xiaowei Guo, Wei-Shi Zheng, Feng Zheng, and Xing Sun. Devil's in the details: Aligning visual clues for conditional embedding in person re-identification, 2020. 1

[12] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification, 2018. 1

[13] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1