Locally Aware Transformer++

Shubham Mittal
ee1180957@iitd.ac.in

Aditi Khandelwal
ee1180434@iitd.ac.in

Abstract

In computer vision-based surveillance, person re-identification (Re-ID) is a well-known issue. The goal of Re-ID is to recognize the same person from numerous non-overlapping viewpoints captured by multiple cameras. Re-ID has piqued the interest of the computer vision community due to the growing demand for intelligent video monitoring. Locally Aware Transformer (LA-TF), inspired from parts-based CNN baselines on Re-ID, aggregates the classification token embeddings into the patch embeddings, and learns an ensemble of $\sqrt{N}$ classifiers, where N is the number of image patches. In this study we present LA-TF++ model that addresses the part-based feature learning problem and improves the performance by 2% mAP on our dataset. Code is available at https://github.com/aditi184/Person_Re-Identification.

1. Introduction

Person re-identification (Re-ID) is the task of identifying a person-of-interest (query person) at other time and/or location captured using same or different camera at same or different orientation. Re-ID is addressed as image-retrieval problem where we have a set of images of different person, called as gallery, and we retrieve the most similar person to the query person from the gallery. Re-ID has wide ranging applications like in surveillance systems that can improve the current situation regarding public safety.

Person Re-ID problem is essentially all about representation learning of the images such that given the a representation or features of the query image and the images in the gallery we can get the most similar person from the gallery using similarity metrics like $L2$, cosine similarity, jaccard's similarity, etc. In this work we do zero-shot inference, i.e., we train a model on a set of person categories and evaluate the model on different set of person categories. Hence, learning robust features that are able to capture the most important features of the image like face, hairstyle, height, skin color so that the same person is identified in different poses and clothing. We may want to avoid features like the color of clothes the person is wearing since the person can appear in different clothing on different days.

CNN-based methods are popular methods to extract robust and discriminative image features but there are two main problems with these methods as discussed in TransReID [2]: (1) CNN-based methods only process one local neighborhood at a time due to a Gaussian distribution of effective receptive fields, and (2) they suffer from loss of fine-level details of the image due to the downsampling operators (like strided and pooling convolution). TransReID [2] uses a pure transformer framework for the object Re-ID task, and outperforms all the CNN-based methods on several popular person/vehicle ReID datasets including MSMT17, Market-1501, DukeMTMCreID, Occluded-Duke, VeRi-776 and VehicleID. Recently Locally Aware Transformer (LA-TF) [5] has shown improvements over TransReID, and thus, we use LA-TF as our baseline in this work. We present our proposed model, LA-TF++ in Section 3.2 that outperforms LA-TF [5] on our dataset, and is also faster (during training and inference) and occupies less storage memory.

2. Related Work

Most of the existing works on Person Re-Identification have been using CNNs [8]. Ye *et al*. [8] provide a survey of existing work till the end of 2020, and summarize the Person Re-ID system in three main components: (1) Feature Representation Learning, which includes both Global and Local Feature Representations of the image, (2) Deep Metric Learning, which is about the design of the loss function, and (3) Ranking Optimization, which focuses on optimizing the retrieved ranking list.

Feature Representation Learning is an important research area wherein the focus is on learning representations of the data such as images which can be useful for various downstream tasks. In Person Re-ID, the features of the person present in the training data are learned in a classification setting, and during inference or retrieval stage, the features of new person are directly used and compared with the features of the people present in the gallery. Hence, learning robust and meaningful features is of utmost importance in this problem which is also emphasized by [7], [9], [2], [5], in some way or the other. AlignReID [9], TransReID [2],

and LA-TF [5] also focus on learning local features that can improve discriminability ability of the learned features. These local features are learned along with global features, and thus, both of them together define an image of a person in a feature space. The hypothesis of feature learning is that the same persons are located close to each other in the feature space.

Deep Metric Learning: In this learning setting, the loss functions are designed to guide the feature representation learning. The combination of Triplet loss and Identity loss is mostly seen in the existing work that has given better performance, i.e., multi-loss training strategy leads to consistent performance gain [8]. AlignReID [9] uses mutual learning approach where the Identity loss is applied on both the global and local features along with a metric loss (based on local and global distances). TransReID [2], on the other hand, uses only the global features for computing the loss (triplet loss + identity loss) since the Transformer architecture captures the local information. LA-TF [5] computes only the identity loss but on an ensemble of classifiers.

Ranking Optimization also plays a crucial role in getting the best retrieval performance. Re-Ranking [11] has shown performance gains in some works ([9]).

3. Methodology

In this section we first provide an in-depth analysis of the baseline model, LA-TF[1] [5] in subsection 3.1. We discuss its limitations and address them in our proposed model, LA-TF++ in Section 3.2.

3.1. Locally Aware Transformer (Baseline)

LA-TF model uses vision transformer, ViT [1] along with an ensemble of classifiers as shown in Figure 2(a). The model architecture is designed to first capture the local features in form of 196 patch embeddings, and then get Globally Enhanced Local Tokens (GELT) by doing weighted sum with the embedding of the CLS token. Finally, they perform row-wise adaptive average pooling to get row embeddings (14 rows, each represented using 768-dim feature vector). Each row embedding is passed into an independent FC Classifier, and the ensemble is trained using one global identity loss or the cross entropy loss. We make following observations:

1. **Overfitting during training? :** LA-TF gets trained on 40 person categories present in the train set, and achieves 99.8% accuracy on the classification task. Since we are not given validation data (comprising of the same 40 person), we do train-val split of the given train set, and observe that there is no overfitting despite high train-accuracy. Thus, we conclude the LA-TF model (or ViT) doesn't overfit on the given train set and we may avoid train-val split for further experiments. Figure 1 shows the learning curves of LA-TF, and confirms that there is no overfitting.

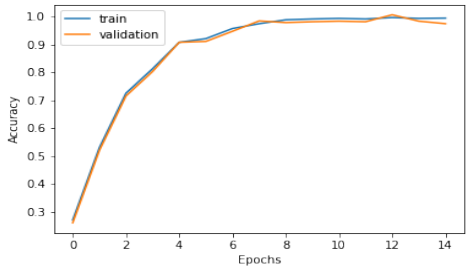


Figure 1: Learning curves of LA-TF

2. **Loss functions during training:** As discussed in Section 2, most works do multi-loss training and use a combination of triplet loss and identity loss. LA-TF uses only the identity loss unlike TransReID and AlignReID.

3. **Local and Global Feature Learning:**

   (a) LA-TF aggregates the global features into the local features, and obtains an Average GELT (L) of shape 14×768. Then it feeds these 14 feature vectors into 14 fully-connected classifiers and learns a combined cross entropy loss.

   (b) TransReID also does the similar thing ,i.e., it applies the loss function on the local features. However, it also applies loss term on the global feature explicitly unlike LA-TF.

Thus, the LA-TF model is not doing good in any of the three main components of Person ReID system as discussed in Section 2:

- It is not doing global feature learning explicitly.
- It doesn't use metric-based loss terms like triplet loss.
- No ranking optimization is performed during retrieval stage.

The retrieval performance of the baseline, LA-TF is shown in Table 2.

3.2. LA-TF++ (Our Model)

We aim to address the limitations of the LA-TF model, as discussed in Section 3.1, and present our improvements below:

[1] code available at https://github.com/SiddhantKapil/LA-Transformer

1. **Loss Function** In order to guide the learned features, both local and global, we add Triplet Loss[2] term to bring same person images closer and different person images farther in the feature space. Further, since the model shows very high classification accuracy of 99.8% on the train set, we use Label Smoothing[3] along with Cross Entropy Loss.

Loss Function	CMC@rank-1	CMC@rank-5	mAP
CE	92.9	96.4	91.5
CELS	92.9	96.4	91.9
CE + TL	89.2	92.9	91.0
CELS + TL	**96.4**	**96.4**	**92.1**

Table 1: Comparison of different loss function using LA-TF on the ValSet. CE: Cross Entropy, CELS: Cross Entropy with Label Smoothing, TL: Triplet Loss

Table 1 shows the comparison of different loss functions using LA-TF. We use mAP score first to compare the results since mAP considers both precision and recall unlike the CMC@rank-k metric (as discussed in [4]). Thus, we conclude from Table 1 that CE+LS is optimal, and it is expected since the LA-TF shows high confidence during its training stage.

2. **Global Feature Learning**: In order to get global feature representation of the image, we apply mean pooling of the Averaged GELT, L (or the row features intuitively) in the Locally Aware Network of LA-TF. This gives one global feature vector, and this is fed into one Fully-Connected Classifier. Our proposed model, LA-TF++ has the same LA-TF backbone with improvements in the Locally Aware network as shown in Figure 2. Hence, LA-TF++ is a smaller and faster model than LA-TF.

3. **Ranking Optimization**: We add re-ranking optimization during the retrieval stage as discussed in [11], but experimentally we found that all the scores drop. Thus, we avoid it but our hypothesis is that it will give performance boost on bigger datasets like Market-1501. Since we have only 28 query images in the validation set, it is difficult to estimate performance boosts. We also added FlipReID [4] but observed no significant improvements (as there is little scope for improvement on small val set), and we omit it in this work.

The retrieval performance of our model, LA-TF++ is shown in Table 2.

[2]Adapted from https://github.com/michuanhaohao/AlignedReID
[3]Adapted from https://github.com/michuanhaohao/AlignedReID

4. Results

Table 2 shows the comparison of our model LA-TF++[4] with the baseline model, LA-TF[5]. We can observe a boost of 1.7% in the mAP and 3.6% in CMC@rank-5 after addressing the limitations of LA-TF.

Model	CMC@rank-1	CMC@rank-5	mAP
Baseline	92.9	96.4	91.5
LA-TF++	**92.9**	**100.0**	**93.2**

Table 2: Retrieval Scores of the two models on validation set

Figure 3 shows the top 10 predictions of both the models on the same query image.

Figure 3: An example query image where LA-TF++ improves upon LA-TF

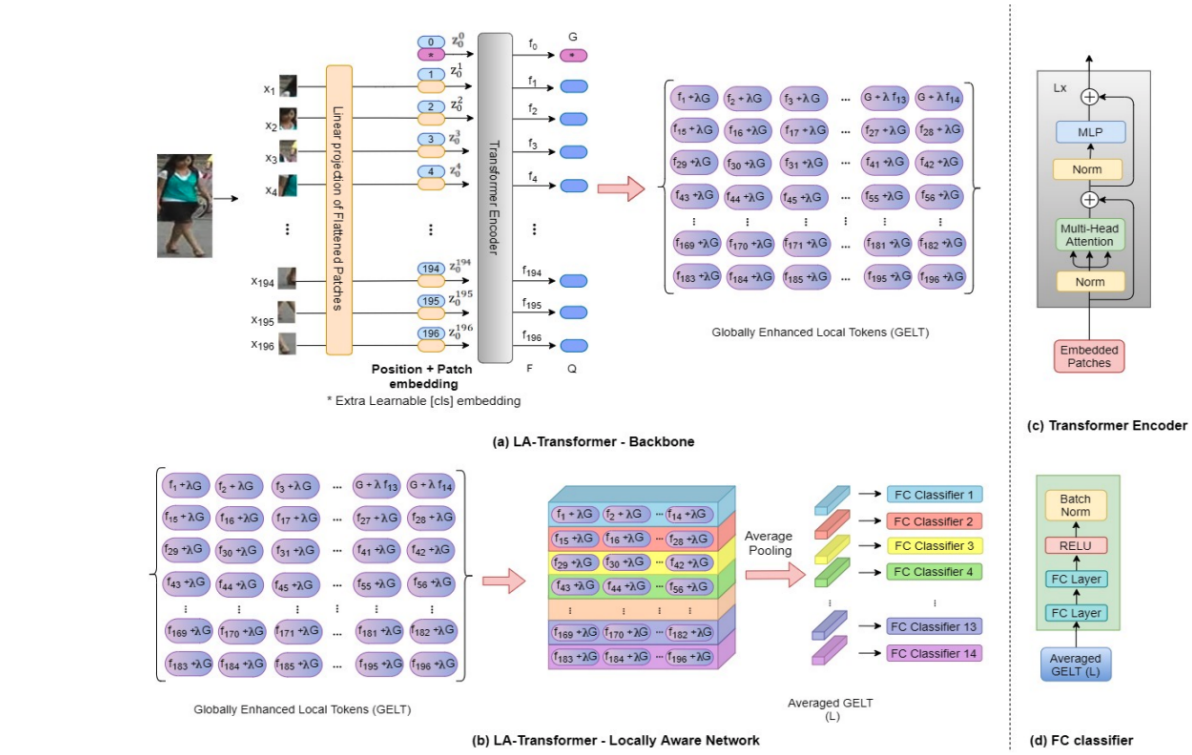
5. Experimental details

Dataset: The dataset (shown in Figure 4) has 114 unique persons, but we use only the train and val set that contains 62 and 12 persons, respectively. Each person has been captured using 2 cameras from 8 different angles.

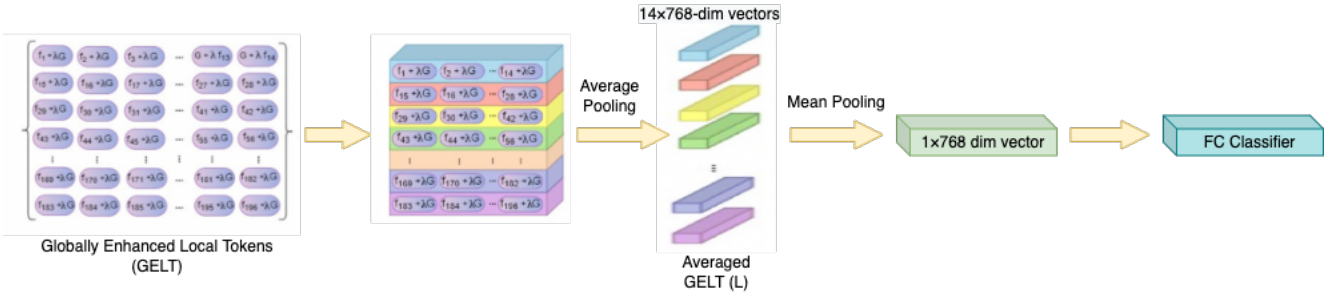
Figure 4: Images of two person (first two, and last two) from our dataset

Locally Aware Transformer: LA-TF has ViT [1] in its backbone which is pre-trained on ImageNet-21K. The im-

[4]model weights available here
[5]baseline model weights available here



(a) Locally Aware Transformer (LA-TF)



(b) LA-TF++ (our model)

Figure 2: (a) shows the LA-TF architecture (backbone and locally aware network). L is the row-wise mean of GELTs obtained from average pooling [5]. (b) shows our proposed model, LA-TF++ architecture where we take the mean of the L and then pass it on to fully-connected classifier

age resolution of our dataset is 48×128 which is warped to 224×224 before passing to ViT. We train the models for 30 epochs with a batch size of 32, an initial learning rate of 3e-4, step decay of 0.8 in Adam optimizer, and $\lambda = 0.8$ (the coefficient in the weighted sum of CLS token into local features). Training and testing of the models were done on Tesla V100 (32 GB) GPU.

Feature vectors similarity search: We use the FAISS library by Johnson *et al*. [3] to calculate the inner product between the query image feature vector and all the gallery images' feature vectors.

Evaluation Metrics: We utilized the widely used metrics Cumulative Matching Characteristics (CMC) [6] and mean Average Precision (mAP) [10] to evaluate the models. CMC@rank-k (also known as Rank-k matching accuracy) determines whether a correct match emerges in the top-k ranked retrieved results or not. Since in an image-retrieval system, there exists multiple ground truths (in the gallery), CMC@rank-k is not a good metric as compared to mAP. mAP is preferred over CMC@rank-k as it considers both, recall and precision.

6. Conclusion

In this work we discussed the limitations of Locally Aware Transformer [5], addressed those problems, and proposed our model, LA-TF++ that outperforms LA-TF on our dataset. LA-TF++ replaces the ensemble of classifiers with a single classifier which not only improves the retrieval performance ,i.e., does better in feature representation learning but also improves the efficiency in terms of space and time.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2, 3

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

[3] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus, 2017. 4

[4] Xingyang Ni and Esa Rahtu. Flipreid: Closing the gap between training and inference in person re-identification, 2021. 3, 5

[5] Charu Sharma, Siddhant R. Kapil, and David Chapman. Person re-identification with a locally aware transformer, 2021. 1, 2, 4, 5

[6] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. Shape and appearance context modeling. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. 4

[7] 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

[8] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook, 2021. 1, 2

[9] 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, 2

[10] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 4

[11] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding, 2017. 2, 3

7. APPENDIX

7.1. Contributions

Both the authors have made equal contributions for making this course project a success. However, for the purpose of project evaluations, following is the break up:

1. Baseline setup: Training and testing script by Shubham and Aditi respectively.

2. Improvements: We did brainstorming together, and thus, we can't break our contributions here. Experimental work and implementation wise: Table 1 (loss functions) by Shubham, and Global Feature Learning by Aditi.

3. We also experimented multiple ideas and added modules from other works (like FlipReID [4], maskRCNN as a background subtraction module) that didn't work and we didn't mention those results.