Person Re-Identification with a Locally Aware Transformer

Shivangi Bithel, Sagar Sharma
Department of Computer Science
IIT Delhi

Abstract

Person re-Identification is a well-known problem in computer-based surveillance systems. The Re-Id aims to identify the same Person in a variety of non-conflicting views from multiple cameras. Due to the growing demand for intelligent video surveillance, Re-ID has gained much interest in the computer vision community. Most person re-Id and object recognition are based on Convolutional Neural Networks (CNNs), but Vision Transformers has begun switching CNNs to perform various computer vision tasks. The main output of the vision transformer is a global classification token, but the vision transformers also produce local tokens that contain additional information about the local regions of the image. Strategies for using these local tokens to improve class accuracy are an active area of research. We are using a Locally Aware Transformer (LA-Transformer) that employs a Parts-based Convolution Baseline (PCB)-inspired strategy for aggregating globally enhanced local classification tokens into an ensemble of $\sqrt{N}$ classifiers, where N is the number of patches. We propose a large 24-block ViT backbone and new data augmentation techniques for robust training as an improvement. We achieves Rank-1 accuracy of 97.7% in our Improvement1 model, 96.6% in our Improvement2 model and 82.6% in our baseline model on the MARKET-1501 dataset. Link to Model and Code

1. Introduction

Transformer models have shown praiseworthy performance on a broad scope of language tasks, e.g. text characterization, machine translation, information retrieval and question answering. The significant effect of Transformer models has become all the more precise with their scalability to huge capacity models. The forward leaps from Transformer networks in Natural Language Processing(NLP) space have started an incredible interest in the computer vision local community to adapt these models for vision and multi-model learning undertakings.

Transformer architectures are based on a self-attention mechanism that learns the relationships between elements of a sequence. Unlike recurrent networks that process sequence elements recursively and can only attend to short-term context, Transformers can attend to complete sequences, thereby learning long-range relationships. Although attention models have been extensively used in both feed-forward and recurrent networks, Transformers are based solely on the attention mechanism and have a unique implementation (i.e., multi-head attention) optimized for parallelization. An essential feature of these models is their scalability to high-complexity models and large-scale datasets, e.g., compared to some of the other alternatives such as intricate attention, which is stochastic and requires Monte Carlo sampling for sampling attention locations. Since Transformers assume minimal prior knowledge about the structure of the problem compared to their convolutional and recurrent counterparts, they are typically pre-trained using pretext tasks on large-scale (unlabelled) datasets. This pre-training avoids costly manual annotations, thereby encoding highly expressive and generalizable representations that model rich relationships between the entities present in a given dataset. The learned representations are then fine-tuned on the downstream tasks in a supervised manner to obtain favorable results.

We work with Vision Transformers for person re-ID and achieved results comparable to the current state-of-the-art CNN-based models. Our approach extends [5] in several ways but primarily because we aggregate the globally enhanced local tokens using a PCB-like strategy that takes advantage of the spatial locality of these tokens. Although [5] use fine-grained local tokens, it does so with a ShuffleNet [19] like Jigsaw shuffling step, which does not take advantage of the 2D spatial locality information inherent in the ordering of the local tokens. LA-Transformer overcomes this limitation by using a PCB-like strategy to combine the globally enhanced local tokens while first preserving their ordering in correspondence with the image dimension. We also adopted blockwise fine-tuning, which can further improve the classification accuracy of LA-Transformer for person re-ID. Blockwise fine-tuning is viable as a form of regularization when training models with many parameters over

relatively small in-domain datasets. [7] advocate blockwise fine-tuning or gradual unfreezing, particularly when training language models due to many fully connected layers. As vision transformers also have high connectivity, we find that this approach can improve the classification accuracy for the LA-Transformer.

This paper is organized as follows: Firstly, we discuss related work involving Transformer architectures and other related methodologies in person re-ID. Secondly, we describe the architecture of LA-Transformer, including locally aware network and blockwise fine-tuning techniques. Finally, we present quantitative results of the person re-ID including mAP and rank-1 analysis on datasets.

2. Related works

For a long time, CNN-based models have dominated image recognition tasks, including person re-ID. Extensive research has been conducted to identify the best strategy for capturing features using CNN to address blurring, background disturbances, partial occlusion, body misalignment, viewpoint changes and pose variation. A pose sensitive embedding to include information about a person's pose in the model, [17] Using the Graph Convolution Network [9] to create a conditional feature vector based on the local correlation between image pairs, [6] Global channel-based and part-based features used, [18] Global pooling was used to capture global features and horizontal pooling after 1×1 CNN for local features. CNN-based methods have made many advances in recent years and are being developed for person re-ID.

Another branch of techniques for person re-ID focuses on developing highly engineered network designs that incorporate additional domain knowledge to enhance re-ID performance. The Part-Aware Approach models function with the primary function and an auxiliary function for each body part. [15] Introduced the idea of calculating part losses by [11] (Part-Based Convulsive Backbone aka PCB) Improved. Even the current best performance models like the Yao and others.PCBs used with domain-specific Spatiotemporal distribution information To get better results in different datasets. In our work, we combine local classifications such as PCB with Vision Transformers, and in addition, we find that our model works better if we pass on global information along with local features. Interest in Vision Transformers initially grew out of attention mechanisms, which were previously used for language translation issues in the NLP, and have significantly impacted image recognition. Introduced parameter-free spatial attention to integrating spatial relationships into global mean pooling (GAP). The Spatial Attention Module (SAM) and the Channel Attention Module (CAM) are used to provide critical spatial and channel information. [3] Position Attention Module (PAM) proposal for semantic pixels in a spatial domain with CAM. Attention mechanisms remain an active area of research for many issues related to object detection and recognition.

Transformers were first introduced in NLP problems by [14], and now Transformers are contributing to many new developments in machine learning. [4] introduced transformers to images by treating a 16x16 patch as a word and treating image classification as analogous to text classification. This approach showed promising results on ImageNet, and it was soon adopted in many image classification problems [10]. Object detection is another highly related problem for which vision transformers have been recently applied [1]. [1] described a correspondence between local tokens and input patches and combined local tokens to create spatial feature maps. At present, this observation of the correspondence between local tokens and input patches has yet to be applied to a wide variety of computer vision problems, nor has it been previously explored in the context of person re-ID. One exception is in the area of image segmentation, for which recent works are beginning to take advantage of the 2D ordering of the local tokens in order to produce more accurate predicted masks [16]. Our approach builds upon the recent work of [5], who was the first to apply vision transformers to object and person re-ID. Although He [5] use global and local tokens, [5] combine the local tokens using a jigsaw classification branch that shuffles the ordering of the local features. Shuffling the order of local features does not take advantage of the observation of [1] in that local features correspond strongly with input patches and therefore have a natural ordering in the form of a 2D image grid. Conversely, LA-Transformer takes advantage of the spatial locality of these local features by combining globally enhanced local tokens with a PCB-like strategy [11]. Furthermore, LA-Transformer incorporates the blockwise fine-tuning strategy described by [7] as a form of regularization for high-connectivity pre-trained language models. As such, LA-Transformer builds upon recent advances in applying vision transformers with novel training techniques to achieve the state of the art accuracy in person re-ID.

3. Method

LA-Transformer combines vision transformers with an ensemble of FC classifiers that take advantage of the 2D spatial locality of the globally enhanced local tokens. Section 3.1 describes the overall architecture including the backbone vision transformer (section 3.1.1), as well as the PCB inspired classifier network ensemble (section 3.1.2). The blockwise fine-tuning strategy is described in section 3.2.

3.1. Locally Aware Transformer

LA-Transformer (figure 1) consists of two main parts: a backbone network and a locally aware network. Both com-

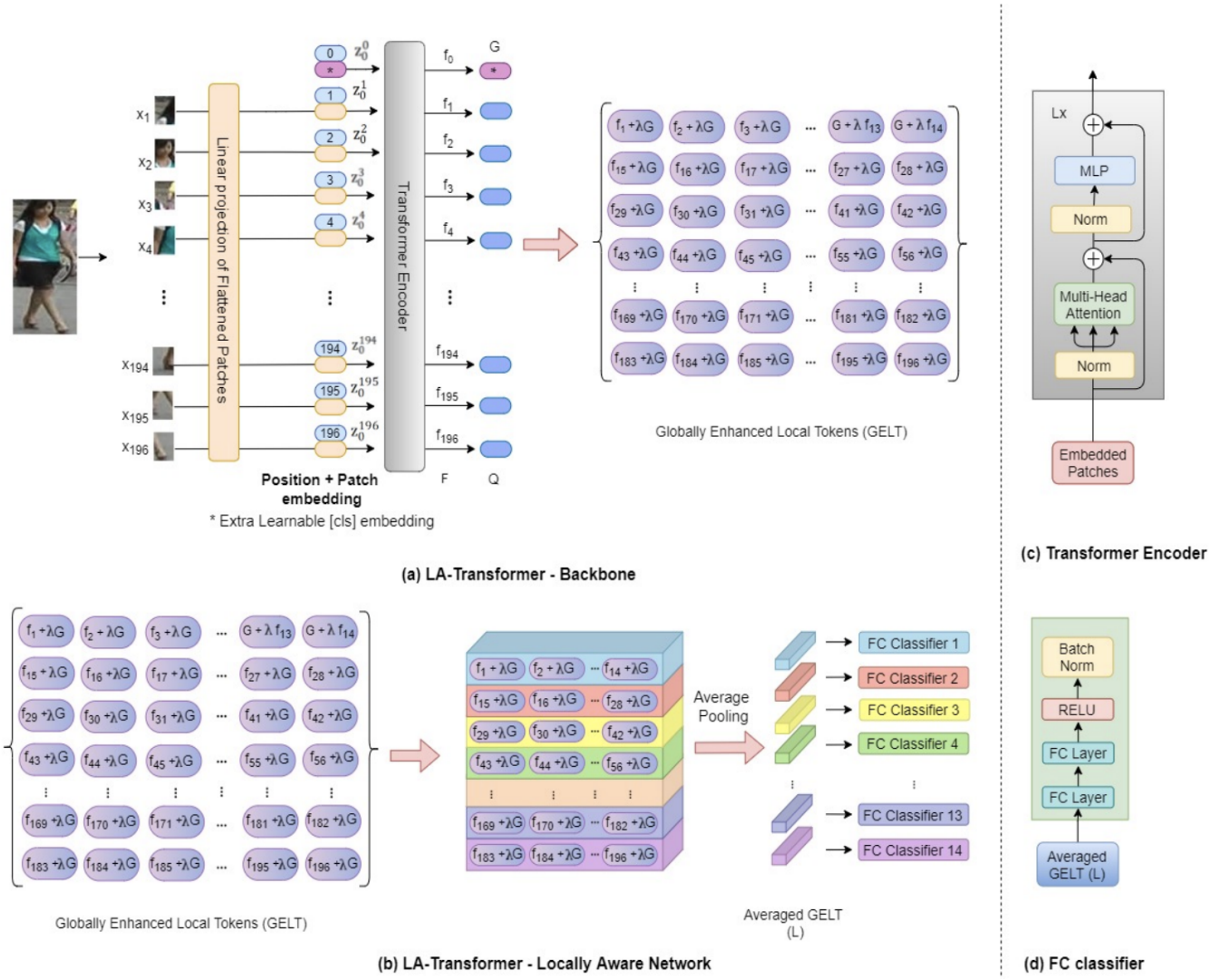


Figure 1. **Outline –** A figure summarizing the entire algorithm

ponents are interconnected and trained as a single neural network model. The backbone network is the ViT architecture as proposed by [4]. ViT generates tokens F = f0, f1, .., fN . The token f0, also known as the global classification token and we refer to this token as the global token G. Supplementary outputs f1..f196 are referred to as local tokens which we denote to collectively as Q. Globally Enhanced Local Tokens (GELT) are obtained by combining global tokens and local tokens (G and Q) using weighted averaging and are arranged into a 14 $\times$ 14 2D spatial grid as seen in Figure 3(a). The row-wise averaged GELTs are then fed to the locally aware classification ensemble as seen in Figure 3(b) to classify during the training process and to generate feature embedding (by concatenating L) during the testing process. These steps are described in greater detail in the following sections 3.1.1 and 3.1.2

3.1.1 LA-Transformer Backbone

The backbone network of LA-Transformer is the ViT (vision transformer). ViT requires extensive training data on the order of 14M to 300M images to train effectively, but the training dataset given to us is relatively small in comparison on the order of 10's of thousands of images. As such we employed a pre-trained ViT model, and further made use of blockwise fine-tuning to improve accuracy as described in section 3.2

Embeddings The backbone ViT architecture takes images of size 224 × 224 as input, and as such the Market1501 and CUHK-03 images are re-sampled to this resolution during training. First, the image is converted into N number of patches Each patch is then linearly projected into D dimensions using the patch embedding function. $x_p^i | i = 1, ...N$.Each patch is then linearly projected into D dimensions using the patch embedding function (E(x_p^i)$|i = 1, ...N$ (eq. 2), which is obtained using a convolution layer with a kernel size of 16 × 16. For non-overlapping patches, a stride equal to 16 is used. D is the number of channels and is set to 768 which represents the size of the embedding. The total number of patches N depends on kernel size, stride, padding, and size of the image. N can be easily calculated using the eq. 1. Assuming padding is 0, and H, W are height and width of an image, KH, KW are height and width of the kernel and S is kernel stride.

$$N = (\frac{H - K_H}{S} + 1) \times (\frac{W - K_W}{S} + 1) \quad (1)$$

Afterward, the learnable class embedding xclass is prepended with the patch embedding (E(x_p^i)) whose output state keeps the information of the entire image and serves as the global vector. The resulting vectors are then added with position embeddings P to preserve the positional information. Subsequently, the final sequence of vectors z0 (eq. 2) is fed into the transformer encoder (figure3) to generate N + 1 feature vectors where N is the number of patches plus class embedding.

$$N = [xclass; E(x_p^1); E(x_p^2)...; E(x_p^N)] + P \quad (2)$$

Transformer Encoder The transformer encoder consist of total B = 12 blocks. Each block contains alternating MSA (Multiheaded Self-Attention) introduced by [10] and MLP blocks. The Layernorm (LN) is applied before MSA and MLP blocks and a residual connections is applied after each encoder block. The output of transformer encoder F described in eq. 5 passes through all the B blocks (eq. 3 and 4).

$$z^{'}{}_b = z_{b-1} + MSA(LN(z_{b-1})) \quad (3)$$

$$z_b = z^{'}{}_b + MLP(LN(z_b)) \quad (4)$$

$$F = LN(z_B) \quad (5)$$

While the seminal work of [4] only uses classification token z_B for classification, LA-Transformer makes use of all of the features z_B eq. 5. Though the class embedding can be removed from the backbone network, our experiments show promising results with class embedding serving as a global vector (Table 2). From our experiments, it is clear that ViT as a backbone network is a good choice for person re-ID based problems. Further, we believe that any transformer based model like Diet by [13], or DeepViT by [18] can be used as a backbone network.

3.1.2 Locally Aware Network

The Locally Aware Network is a classifier ensemble similar to the PCB technique of [11] but with some differences. Firstly, in [11] the input features are purely local, whereas in LA-Transformer, we find that the inclusion of global vectors along with local vectors via weighted averaging can increase the network accuracy. Secondly, although in [11] the image is divided into six input regions, we divide the 2D spatial grid of tokens into $\sqrt{N}$= 14 regions as seen in Figure 3. Finally, while PCB uses a convolutional backbone, LA-Transformer uses the ViT backbone.

In Figure 3, the transformer encoder outputs N + 1 feature vectors. The global tokens G = f_0 and local tokens Q = [f_1,f_2,f_3, ...,f_N] are obtained for which N is number of patches. N_R is defined as the total number of patches per row and N_C as the total number of patches per column. In our case, N_R=N_C =$\sqrt{N}$.Then we define L as the averaged GELT obtained after average pooling of Q and G as follows,

$$L_i = \frac{1}{N_R} \cdot \sum_{j=i*N_R+1}^{(i+1)*N_R} \frac{Q_j + \lambda G}{(1+\lambda)} i = 0...N_{C-1} \quad (6)$$

In eq. 6 all the patches in a row are averaged to create one local vector per row . The total number of FC classifiers is equal to N_C . Each FC classifier contains two fully connected layers with RELU and Batch Normalization. We define y as the output of LA-Transformer as follows,

$$y_i = FC_i(L_i) \quad (7)$$

The outputs y are passed through softmax and the softmax scores are summed together. The argument of the maximum score represents the ID of the person as follows.

$$score = \sum_{i=0}^{N_C} softmax(y_i) \quad (8)$$

$$prediction = argmax(score) \quad (9)$$

3.2. Fine-tuning Strategies

According to the recent studies of [13] and [4], training a vision transformer from scratch requires about 14M-300M images. Person re-ID datasets are known or their small size and training a transformer on these datasets can quickly lead to overfitting. As such, ViT was pre-trained on ImageNet, and then fine-tuned on person re-ID datasets. Blockwise fine-tuning was applied which is highly similar to the gradual unfreezing method described by [7] for the purposes of training large language models in the event of limited training data from a target domain.

Blockwise Fine-tuning In blockwise fine-tuning, all transformer blocks are frozen in the start except for the bottleneck model. After every t epochs (where t is a hyperparameter), one additional transformer encoder block is unfrozen and the learning rate is reduced as described by Algo 3.2. Blockwise fine-tuning helps in mitigating the risk of catastrophic forgetting of the pre-trained weights [7]. The learning rate decay helps in reducing the gradient flow in the subsequent layers hence prevent abrupt weight updates.

Algorithm 1 Blockwise Fine-tuning

Freeze all the transformer blocks B
Initialize parameters:
t = 2, b = 12, lr = 3e-4,lr-decay=0.85
while $i \geq 0$ **and** $i < epochs$ **do**
  if $i \mod t == 0$ & $b > 0$ **then**
    unfreeze B[b]
    $b \leftarrow b - 1$
    $lr \leftarrow lr * lr_decay$
  end if
end while

3.3. Improvement

We propose to change the backbone of ViT-base used to generate image global and local embeddings in LA-Transformer. We tried various models like DeiT [13], MLP Mixer model [12] and ViT Large [4]. We also introduced new data augmentation techniques inspired from "Benchmarks for Corruption Invariant Person Re-identification" [2] like Random horizontal flipping, random vertical flipping and Colour Jitter to make our model robust to corrupted images. To avoid over-fitting of our model we used train set of MARKET-1501 dataset which contains 751 training classes. Table 1 shows the ablation study of various backbones and their validation scores on the 12 class val set given to us. As we can observe that with ViT large, we get 100% Rank-1 accuracy and 99.14% MAP therefore we choose ViT large as our improved backbone model.

We are proposing two models using the ViT-Large backbone architecture. Our Improvement1 model uses ViT-L, which takes input, an image of 384 x 384 size, and divides it into patches of size 16 x 16. It generates 567 such patches per image, and the output size of each patch embedding is 1 x 1024. We also tried with 32 x 32 size patches, but the accuracy decreased, which means that 16 x 16 size patches capture local tokens better than a 32 x 32 patch. Our Improvement2 model takes input, an image of size 224 x 224, and generates 196 patches of size 16 x 16, similar to baseline, but the output embedding dimension of each patch is 1 x 1024. Our baseline LA-transformer takes input, an image of 224 x 224, and generates 196 patches of 16 x 16. The output embedding size of our baseline model is 1 x 768.

Backbone model	**Rank-1**	**Rank-5**	**MAP**
ViT-B/16 224 (baseline)	0.9285	1	0.89
DeiT-B/16_distilled 224	0.8571	0.9285	0.7101
DeiT-B/16 224	0.8214	0.8928	0.7942
ViT-L/16 224	**1**	**1**	**0.9914**

Table 1. Validation Results for different backbone models

The ViT-L/16 model uses 24 transformer blocks and thus trains more parameters. The large output embedding size is capable of capturing more information.

4. Results

4.1. Datasets and Metrics

- The dataset has 114 unique persons. The train and val set contain 62 and 12 persons, respectively.
- Each person has been captured using 2 cameras from 8 different angles. That is, each person would have 16 images. All images of a unique person is stored in a single directory (numbered from 001 to 114).
- The images of a person in the val set in split into query and gallery images. The query is the set of images which will be used to retrieve the images of the same person from the gallery.
- Query and gallery are mutually exclusive sets.

Re-ID is evaluated using Cumulative Matching Characteristics (CMC) and Mean Average Precision (mAP). We are using the below metrics to evaluate our models.

- **Rank-1**:Number of relevant gallery images out of total relevant images in the top 1 ranked results, averaged over all queries.
- **Rank-5**:Number of relevant gallery images out of total relevant images in the top 5 ranked results, averaged over all queries.
- **MAP**:It is the mean of the Average Precision of a query over all the queries that are considered.

4.2. Model Implementation Details

Implementation details like backbone model, input image size, output embedding size, patch size, number of patches, number of transformer blocks, batch size, epochs, learning rate, blockwise unfreezing counter, learning rate decay and number of training classes of baseline, improvement1 and improvement2 model is listed in Table 2.
Baseline: ViT-B/16 was pre-trained on large image dataset and used as a backbone network for baseline model. All the

Model	Baseline	Imp1	Imp2
Backbone	ViT-B/16	ViT-L/16	ViT-L/16
Input image size	224 x 224	384 x 384	224 x 224
Patch size	16	16	16
Number of patches	196	576	196
Transformer blocks	12	24	24
Output dimension	1 x 768	1 x 1024	1 x 1024
Batch size	32	8	32
Epochs	30	15	30
Learning rate	3e-5	2e-5	2e-5
t	2	1	2
Learning rate decay	0.8	0.8	0.8
Training classes	62	825	62

Table 2. Implementation Details of baseline and proposed models

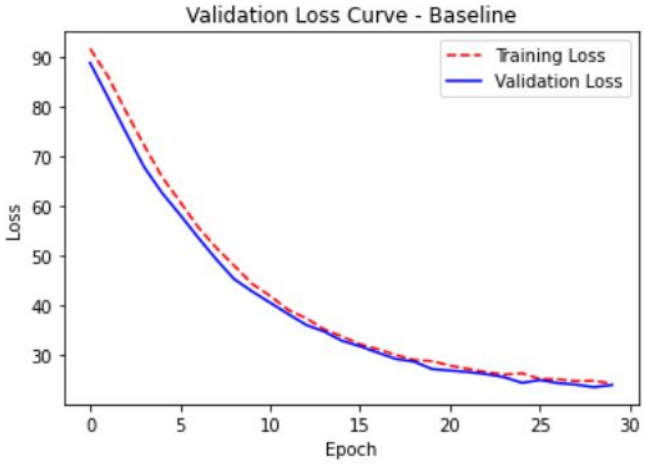
images are resized into 224 $\times$ 224 as this resolution is compatible with the backbone network. The model is trained over 30 epochs with a batch size of 32. We used the Adam optimizer with an initial learning rate of 3e $-$ 5, step decay of 0.8, t=2, b=12 and lamda is 0.8. For testing, we concatenated all of the averaged GELTs L to generate the feature embedding. To efficiently calculate the Euclidean norm between the query and gallery vectors, we use the FAISS library [8] .

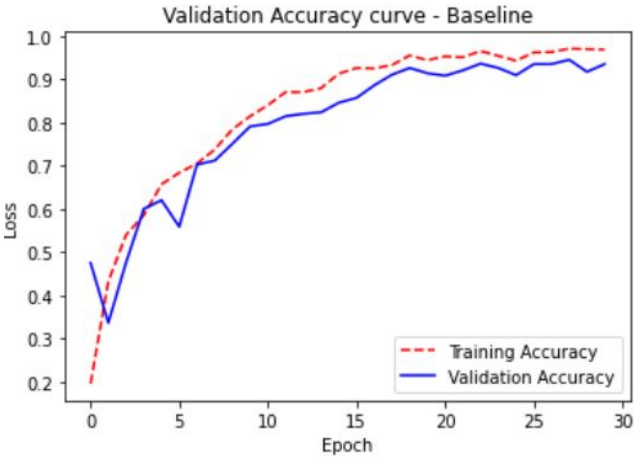
Improvement1: ViT-L/16 was pre-trained on large image dataset and used as a backbone network for Improvement1 model. All the images are resized into 384 $\times$ 384 as this resolution is compatible with the backbone network. The model is trained over 15 epochs with a batch size of 8. We used the Adam optimizer with an initial learning rate of 2e $-$ 5, step decay of 0.8, t=1, b=24 and lamda is 0.8. For testing, we concatenated all of the averaged GELTs L to generate the feature embedding. To efficiently calculate the Euclidean norm between the query and gallery vectors, we use the FAISS library.

Improvement2: ViT-L/16 was pre-trained on large image dataset and used as a backbone network for Improvement2 model. All the images are resized into 224 $\times$ 224 as this resolution is compatible with the backbone network. The model is trained over 15 epochs with a batch size of 32. We used the Adam optimizer with an initial learning rate of 2e $-$ 5, step decay of 0.8, t=2, b=24 and lamda is 0.8. For testing, we concatenated all of the averaged GELTs L to generate the feature embedding. To efficiently calculate the Euclidean norm between the query and gallery vectors, we use the FAISS library.

5. Conclusions and Future Work

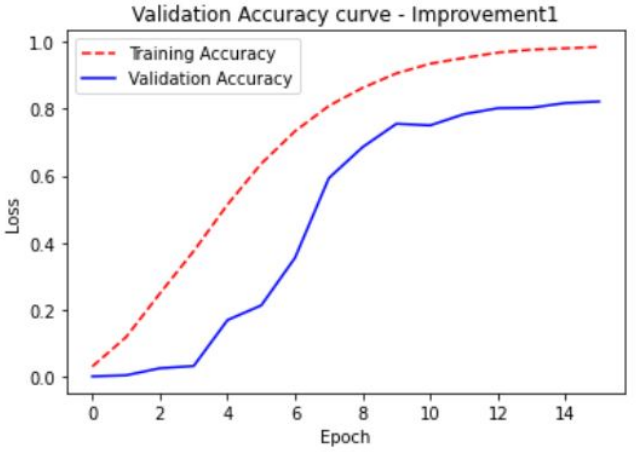
We use re-id technique called Locally Aware Transformer (LA-Transformer) which achieves great performance on given dataset. In this approach firstly, we observed the effect of global token and local token of vision transformer in combination (with a PCB-like strategy). Secondly, we observed the effectiveness of block-wise fine-tuning to regularize the fine-tuning of a pre-trained vision transformer backbone network. Our proposed Improvement with ViT-L and data augmentation techniques can extract rich structured patterns even if the images are corrupted. In future we will explore more data augmentation techniques like Random erasing, etc to improve the robustness of our model. We would also like to add cross modality meta data like text, describing the query image to further enhance the performance of our model.

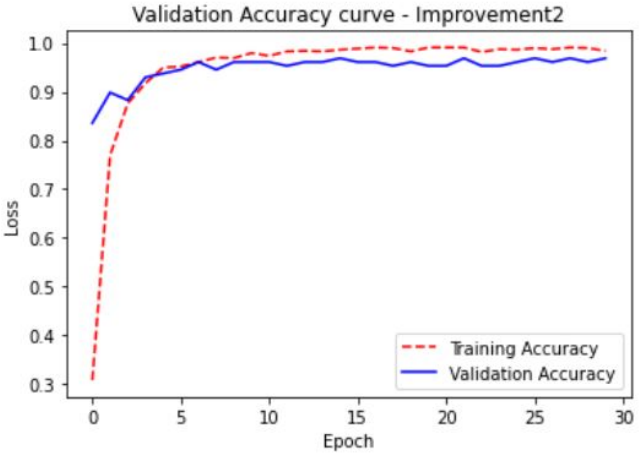


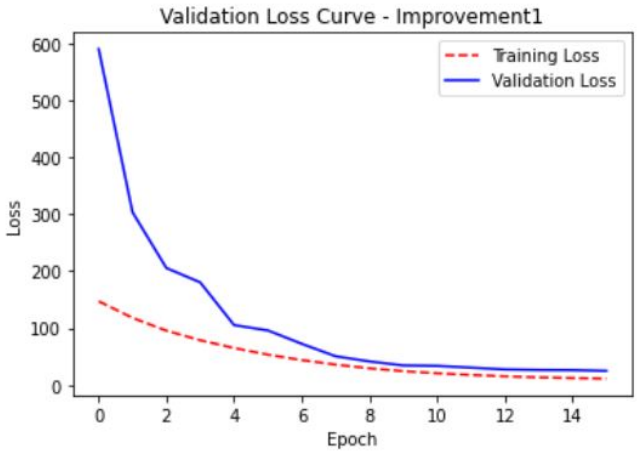


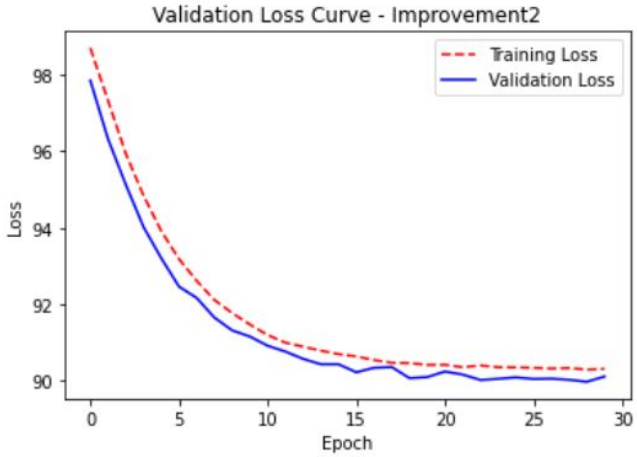
6. Acknowledgement

This project was given to us as a coursework for COL780. We are grateful to our professor Dr. Chetan Arora and our TA Mr. Soumen Basu for providing us this opportunity to work on this challenging task and guiding us throughout the project.









References

[1] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Toward transformer-based object detection, 2020. 2

[2] Minghui Chen, Zhiqiang Wang, and Feng Zheng. Benchmarks for corruption invariant person re-identification, 2021. 5

[3] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification, 2019. 2

[4] 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, 4, 5

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

[6] Fabian Herzog, Xunbo Ji, Torben Teepe, Stefan Hörmann, Johannes Gilg, and Gerhard Rigoll. Lightweight multi-branch network for person re-identification, 2021. 2

[7] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018. 2, 4, 5

[8] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus, 2017. 6

[9] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017. 2

[10] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer, 2018. 2, 4

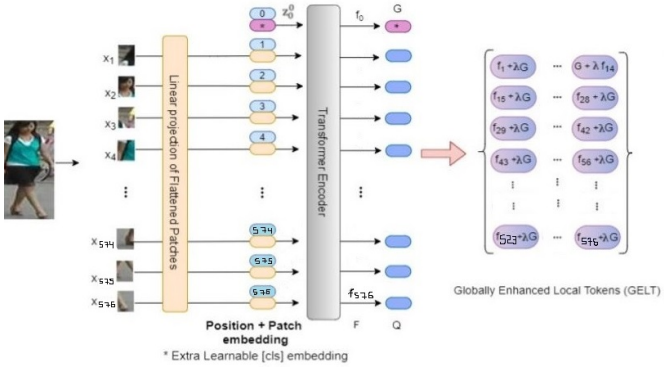


Figure 2. **Backbone Architecture of Improvement1 model**

[11] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), 2018. 2, 4

[12] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision, 2021. 5

[13] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention, 2021. 4, 5

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

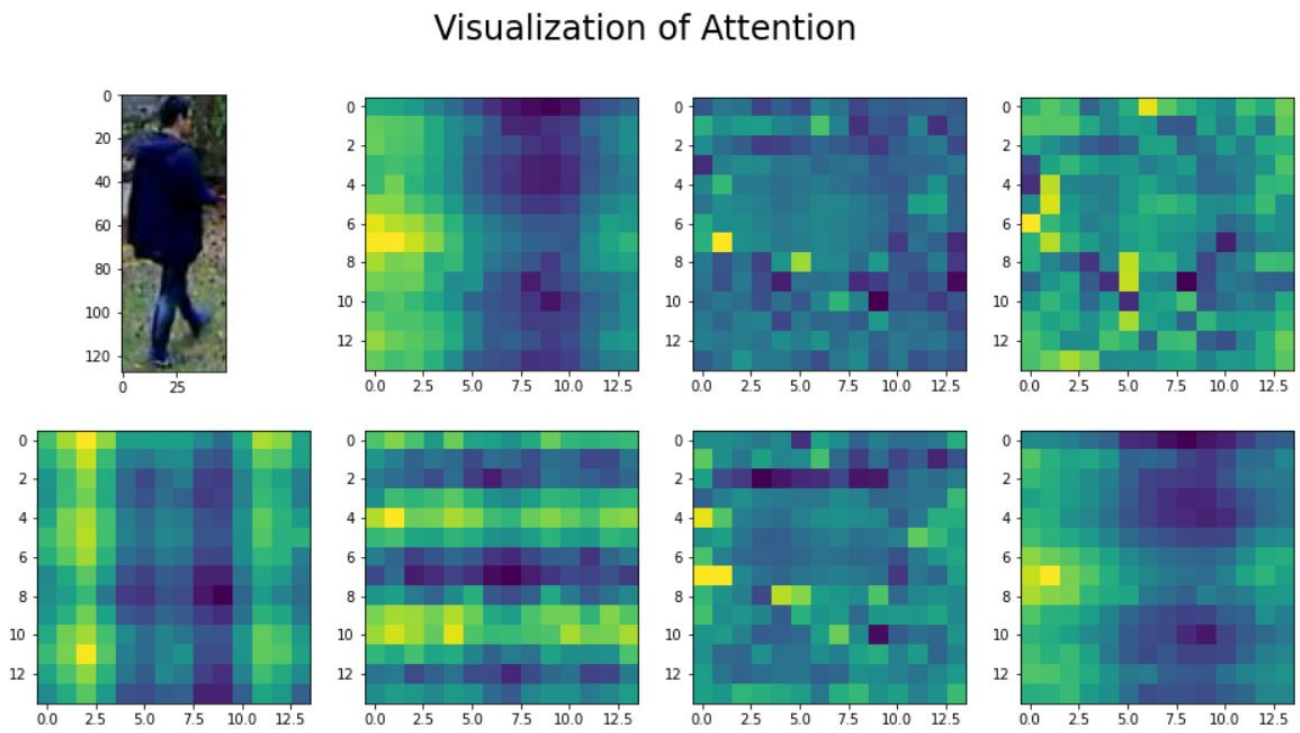
[15] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification, 2018. 2

[16] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020. 2

[17] 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. 2

[18] 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. 2, 4

[19] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices, 2017. 1



A. Ablation Study

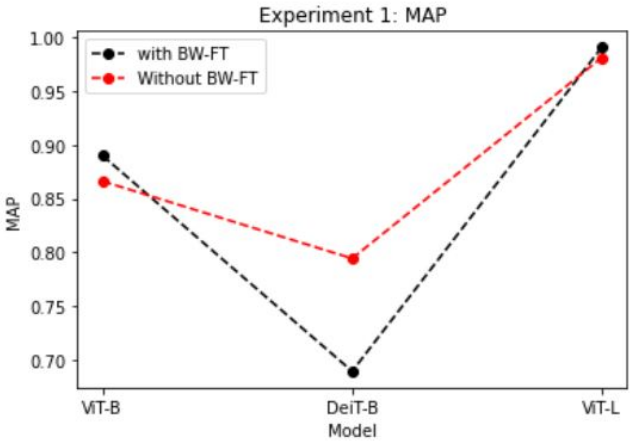
We performed various experiments on our model to understand its robustness in Re-Identification task.

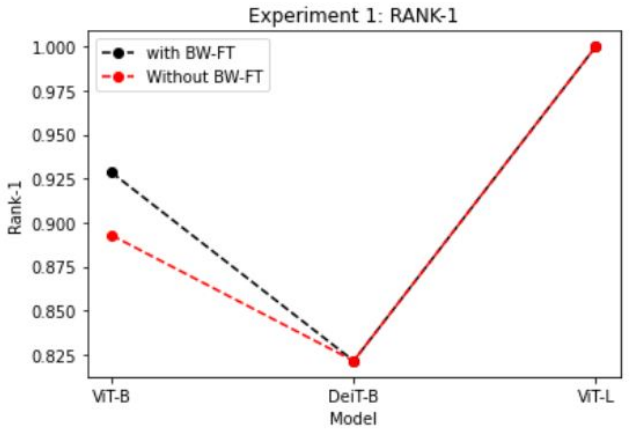
A.1. Experiment-1: Affect of Blockwise fine-tuning

To understand the affect of blockwise finetuning we evaluated three different backbone models ViT-B, DeiT-B and ViT-L on our Valdiation set. The results are mentioned in Table 3. From the results, we can observe that blockwise finetuning is beneficial for ViT model but not for DeiT model.

Model Backbone	With BW-FT		Without BW-FT	
	Rank-1	MAP	Rank-1	MAP
ViT-B/16 224	0.9285	0.89	0.8928	0.8662
DeiT-B/16 224	0.8214	0.6891	0.8214	0.7942
ViT-L/16 224	**1**	**0.991**	1	0.9814

Table 3. Affect of Blockwise fine-tuning



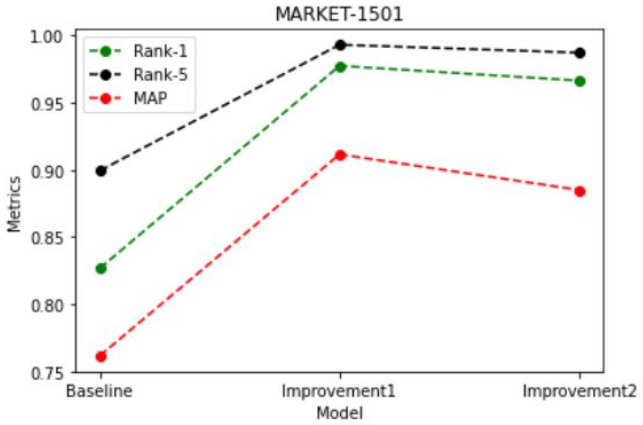


A.2. Experiment-2: Test result on MARKET-1501 dataset

As test set was not available, we used the test set of Market-1501 dataset which has 750 classes and observed the results as mentioned in Table 4. From the results we can see that our proposed model has improved with a significant margin on the baseline model. Our Improvement1 model achieves 97.7% Rank-1 accuracy, 99.3% Rank-5 accuracy and 91.1% MAP on MARKET-1501 dataset.

Model	Rank-1	Rank-5	MAP
Baseline	0.8266	0.8993	0.7615
Improvement1	**0.9774**	**0.9931**	**0.9115**
Improvement2	0.9664	0.9872	0.885

Table 4. Test result on MARKET-1501 dataset



A.3. Experiment-3: MAP vs Rank-1 on Val set

We plotted MAP vs Rank-1 graph on our validation set to observe the relationship between the two metrics across our various backbone models. ViT-L performs better than all the other backbone models.