

AdaptiveReID: Adaptive L_2 Regularization in Person Re-Identification

Xingyang Ni, Liang Fang, Heikki Huttunen

Faculty of Information Technology and Communication Sciences

Tampere University, Finland

Email: {xingyang.ni, liang.fang, heikki.huttunen}@tuni.fi

Abstract—We introduce an adaptive L_2 regularization mechanism termed AdaptiveReID, in the setting of person re-identification. In the literature, it is common practice to utilize hand-picked regularization factors which remain constant throughout the training procedure. Unlike existing approaches, the regularization factors in our proposed method are updated adaptively through backpropagation. This is achieved by incorporating trainable scalar variables as the regularization factors, which are further fed into a scaled hard sigmoid function. Extensive experiments on the Market-1501, DukeMTMC-reID and MSMT17 datasets validate the effectiveness of our framework. Most notably, we obtain state-of-the-art performance on MSMT17, which is the largest dataset for person re-identification. Source code will be published at <https://github.com/mixingyang/AdaptiveReID>.

I. INTRODUCTION

Person re-identification involves retrieving corresponding samples from a gallery set based on the appearance of a query sample across multiple cameras. It is a challenging task since images may differ significantly due to variations in factors such as illumination, camera angle and human pose. On account of the availability of large-scale datasets [1], [2], [3], remarkable progress has been witnessed in recent studies on person re-identification, *e.g.*, utilizing local feature representations [4], [5], leveraging extra attribute labels [6], [7], improving policies for data augmentation [8], [9], adding a separate re-ranking step [10], [11] and switching to video-based datasets [12], [13].

L_2 regularization imposes constraints on the parameters of neural networks and adds penalties to the objective function during optimization. It is a commonly adopted technique which can improve the model’s generalization ability. Although some works [14], [15], [16], [17] provide insights on the underlying mechanism of L_2 regularization, it is an understudied topic and has not received sufficient attention. In most literature, L_2 regularization is taken for granted, and the text dedicated to it is typically shrunk into one sentence as in [18]. On the other hand, existing approaches assign constant values to regularization factors in the training procedure, and such hyperparameters are hand-picked via hyperparameter optimization which is a tedious and time-consuming process. The primary purpose of this work is to address the bottleneck of conventional L_2 regularization and introduce a mechanism which learns the regularization factors and update those values adaptively.

In this paper, our major contributions are twofold:

- We introduce an adaptive L_2 regularization mechanism, which optimizes each regularization factor adaptively as the training procedure progresses.
- With the proposed framework, we obtain state-of-the-art performance on MSMT17, which is the largest dataset for person re-identification.

The rest of this paper is organized as follows. Section II reviews important works in person re-identification and L_2 regularization. In Section III, we present the essential components of our baseline, alongside the proposed adaptive L_2 regularization mechanism. Section IV describes the details of our experiments, including datasets, evaluation metrics and comprehensive analysis of our proposed method. Finally, Section V concludes the paper.

II. RELATED WORK

In this section, we give a brief overview of two research topics, namely, person re-identification and L_2 regularization.

A. Person Re-Identification

Utilizing local feature representations which are specific to certain regions, has been shown successful. Varior *et al.* [4] propose a Long Short-Term Memory architecture which models the spatial dependency and thus extracts more discriminative local features. Sun *et al.* [5] apply a uniform partition strategy which divides the feature maps evenly into individual parts, and the part-informed features are concatenated to form the final descriptor.

Besides, methods based on auxiliary feature are advocated, aiming to utilize extra attributes in addition to the identity labels. Su *et al.* [6] shows that learning mid-level human attributes can be used to address the challenge of visual appearance variations. Specifically, an attribute prediction model is trained on an independent dataset which contains the attribute labels. Lin *et al.* [7] manually annotate attribute labels which contain detailed local descriptions. A multi-task network is proposed to learn an embedding for re-identification and also predict the attribute labels. In addition to the performance improvement in re-identification, such system can speed up the retrieval process by ten times.

By applying random manipulations on training samples, data augmentation has played an essential role in suppressing the overfitting issue and improving the generalization of models. Zhong *et al.* [8] introduce an approach which erases the pixel values in a random rectangle region during training. By contrast, Dai *et al.* [9] suggest dropping the same region for all samples in the same batch. Such feature dropping branch strengthens the learned features of local regions.

Adding a separate re-ranking step to refine the initial ranking list can lead to significant improvements. Zhong *et al.* [10] develop a k-reciprocal encoding method based on the hypothesis that a gallery image is more likely to be a true match if it is similar to the probe in the k-reciprocal nearest neighbours. Zhou *et al.* [11] rank the predictions with a specified local metric by exploiting negative samples for each online query, rather than implementing a general global metric for all query probes.

Lastly, some works shift the emphasis from image-based to video-based person re-identification. Liu *et al.* [12] introduce a spatio-temporal body-action model which exploits the periodicity exhibited by a walking person in a video sequence. Alternatively, Dai *et al.* [13] present a learning approach which unifies two modules: one module extracts the features of consecutive frames, and the other module tackles the poor spatial alignment of moving pedestrians.

B. L_2 regularization

Laarhoven [14] prove that L_2 regularization would not regularize properly in the presence of normalization operations, *i.e.*, batch normalization [19] and weight normalization [20]. Instead, L_2 regularization will affect the scale of weights, and therefore it has an influence on the effective learning rate.

Similarly, Hoffer *et al.* [15] investigate how does applying weight decay before batch normalization affect learning dynamics. Combining weight decay and batch normalization would constrain the norm to a small range of values and lead to a more stable step size for the weight direction. It enables better control over the effective step size through the learning rate.

Later on, Loshchilov *et al.* [16] clarify a long-established misunderstanding that L_2 regularization is equivalent to weight decay. The aforementioned statement does not hold when applying adaptive gradient algorithms, *e.g.*, Adam [21]. Furthermore, they suggest decoupling the weight decay from the optimization steps, and it leads to the original formulation of weight decay.

Most recently, Lewkowycz *et al.* [17] present an empirical study on the relations among the L_2 coefficient, the learning rate, and the number of training epochs and the performance of the model. In a similar manner as learning rate schedules, a manually designed schedule for the L_2 parameter is proposed to increase training speed and boost model's performance.

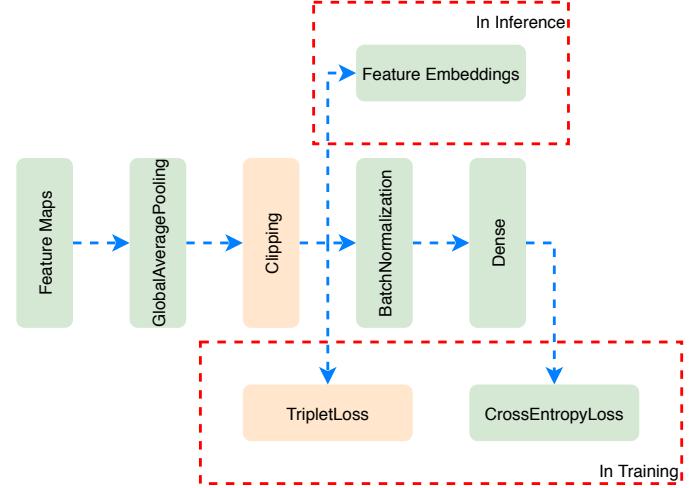


Fig. 1. Structure of an objective module. In the training procedure, two objective functions are applied: triplet loss [22] and categorical cross-entropy loss. In the inference procedure, the feature embeddings before the batch normalization [19] layer are extracted as the representations. Note that blocks in yellow are excluded from the minimal setup.

III. PROPOSED METHOD

In this section, we first present a minimal setup for person re-identification. Later on, we explain five components that contribute to significant improvements in performance and use the resulting method as the baseline in our study. Most importantly, we discuss the proposed adaptive L_2 regularization mechanism at the end.

A. Minimal setup

Backbone: ResNet50 [18], initialized with ImageNet [23] pre-trained weights, is selected as the backbone model. For convenience, it is separated into five individual blocks, *i.e.*, block 1-5, as illustrated in Figure 2. Additionally, the stride arguments of the first convolution layer in block 5 are set to 1, rather than default value 2. This enlarges the feature maps by a scale factor of 2 along with both height and width dimensions while reusing the pre-trained weights and keeping the total amount of parameters identical.

Objective module: Figure 1 demonstrates the structure of an objective module that converts the feature maps to learning objectives. A global average pooling layer squeezes the spatial dimensions in the feature maps, and the following batch normalization [19] layer generates the normalized feature vectors. The concluding fully-connected layer does not contain a bias vector, and it produces the predicted probabilities of each unique identity so that the model can be optimized using the categorical cross-entropy loss. In the inference procedure, the feature embeddings before the batch normalization layer are extracted as the representations, and cosine distance is adopted to measure the distance between two samples.

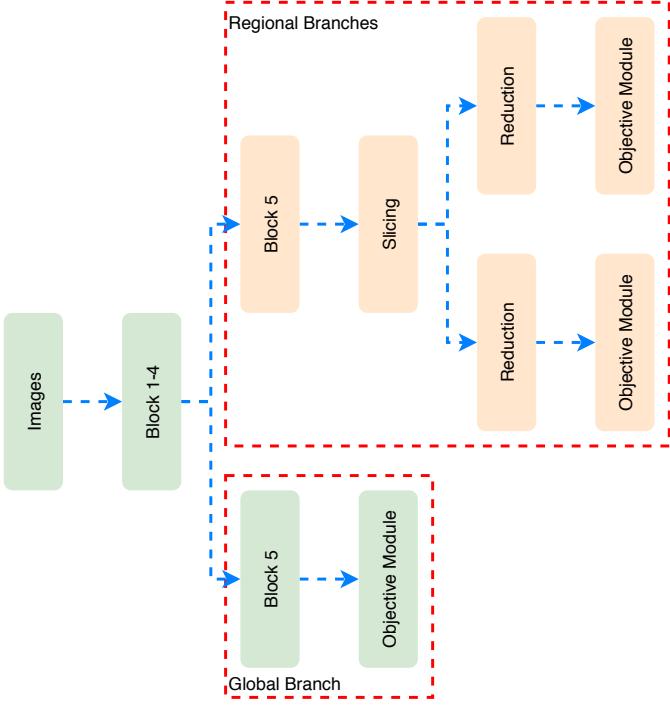


Fig. 2. Topology of the overall model. Feature embeddings from multiple objective modules are concatenated in the inference procedure. Note that blocks in yellow are excluded from the minimal setup.

Overall topology: The topology of the overall model is shown in Figure 2. It is to be observed that the minimal setup only contains the global branch. Given a batch of images, the individual blocks from the backbone model utilize successively, and an objective module is appended at the end.

Data augmentation: The image is resized to target resolution using a bilinear interpolation method. Besides, the image is flipped horizontally at random with probability set to 0.5. Zero paddings are added to all sides of the image, *i.e.*, the top, bottom, left, and right sides. A random part with target resolution is subsequently cropped.

Learning rate: The learning rate increases linearly from a low value to the pre-defined base learning rate in the early stage of the training procedure, and it is divided by ten once the performance on the validation set plateaus. On the one hand, the warmup strategy suppresses the distorted gradient issue at the beginning [24]. On the other hand, periodically reducing the learning rate boosts the performance even further.

Label smoothing: The label smoothing regularization [25] is applied alongside with the categorical cross-entropy loss function. Given a sample with ground truth label $y \in \{1, 2, \dots, N\}$, the one-hot encoded label $\mathbf{q}(i)$ equals to 1 only if the index i is as the same as label y , and 0 otherwise. The smoothed label introduces a hyperparameter $\epsilon \in (0, 1)$ and is calculated as:

$$\mathbf{q}'(i) = (1 - \epsilon)\mathbf{q}(i) + \frac{\epsilon}{N}. \quad (1)$$

B. Baseline

Triplet loss: As highlighted in Figure 1, the triplet loss [22] is applied on the feature embeddings before the batch normalization layer. It mines the moderate hard triplets instead of all possible combinations of triplets, given that using all possible triplets may destabilize the training procedure. Considering that multiple loss functions are present, the weighting coefficients of each loss function are set to 1 on account of simplicity.

Regional branches: In addition to the global branch, two regional branches are integrated into the model. Figure 2 illustrates the diagram of those regional branches. Firstly, the block 5 from the backbone model is replicated, and it is not shared with the global branch. Secondly, we adopt the uniform partition scheme as in [5]. The slicing layer explicitly divides the feature maps into two horizontal stripes. Lastly, dimensionality reduction is performed using a convolutional layer on each stripe. Separate objective modules are appended afterwards. In the inference procedure, feature embeddings from multiple objective modules are concatenated.

Random erasing: In addition to random horizontal flipping, random erasing [8] is utilized in data augmentation. During training, it erases an area of original images to improve the robustness of the model, especially for the occlusion cases.

Clipping: The clipping layer is inserted between the global average pooling layer and the batch normalization layer in Figure 1. It performs element-wise value clipping so that values in its output are contained in a closed interval. The clipping layer works in a similar manner as the ReLU-n units [26], and it relieves optimization difficulties in the succeeding triplet loss [22].

L_2 regularization: Conventional L_2 regularization is utilized to all trainable parameters, *i.e.*, the regularization factors remain constant throughout the training procedure. Additionally, those regularization factors need to be hand-picked via hyperparameter optimization.

C. Adaptive L_2 regularization

A neural network consists of a set of N distinct parameters,

$$P = \{\mathbf{w}_n \mid n = 1, \dots, N\}, \quad (2)$$

with P containing all trainable parameters. Each \mathbf{w}_n is an array which could be a vector, a matrix or a 3rd-order tensor. For example, the kernel and bias terms in a fully-connected layer are a matrix and a vector, respectively.

Conventional L_2 regularization imposes an additional penalty term to the objective function, which can be formulated as follows:

$$L_\lambda(P) = L(P) + \lambda \sum_{n=1}^N \|\mathbf{w}_n\|_2^2, \quad (3)$$

where $L(P)$ and $L_\lambda(P)$ denote the original and updated objective functions, respectively. In our case (see Figures 1 and 2), $L(P)$ is a weighted sum of triplet loss [22] and categorical cross-entropy loss functions. In addition, $\|\mathbf{w}_n\|_2^2$

refers to the square of the L_2 norm¹ of \mathbf{w}_n , and the constant coefficient $\lambda \in \mathbb{R}_+$ defines the regularization strength.

One may wish to add penalties in a different way, *e.g.*, applying lighter regularization in the early layers but stronger in the last ones. Thus, it is possible to generalize even further, *i.e.*, defining a unique coefficient for each $\|\mathbf{w}_n\|_2^2$:

$$L_\lambda(P) = L(P) + \sum_{n=1}^N (\lambda_n \|\mathbf{w}_n\|_2^2), \quad (4)$$

where each parameter \mathbf{w}_n is associated with an individual regularization factor $\lambda_n \in \mathbb{R}_+$.

Obviously, it is infeasible to manually fine-tune those regularization factors λ_n for $n = 1, \dots, N$ one by one, since N is in the order of 100 for models trained with ResNet50. Therefore, we treat them as any other learnable parameters and find suitable values from the data itself.

To make the aforementioned regularization factors adaptive, a straightforward extension is obtained by replacing the pre-defined constant λ_n with scalar variables which are trainable through backpropagation. After the modification, Equation 4 remains unchanged while $\lambda_n \in \mathbb{R}$. However, such an approach without any constraints on λ_n will fail. Namely, setting negative values for λ_n allows naively increasing $\|\mathbf{w}_n\|_2^2$ so that $L_\lambda(P)$ decreases sharply. In other words, the L_2 regularization penalties would become dominant in the optimization process. Thus the model collapses and would not learn useful feature embeddings.

To address the collapse problem, we apply the hard sigmoid function which assures that the regularization factor λ_n would always have non-negative values. The hard sigmoid function is defined as

$$f(x) = \begin{cases} 0, & \text{if } x < -c \\ 1, & \text{if } x > c \\ x/(2c) + 0.5, & \text{otherwise.} \end{cases} \quad (5)$$

In our experiments, we use $c = 2.5$, but any other positive values can be used as well.

The regularization factor λ_n is obtained by applying the hard sigmoid on the raw parameters as

$$\lambda_n = f(\theta_n), \quad (6)$$

where $\theta_n \in \mathbb{R}$ ($n = 1, \dots, N$) are the trainable scalar variables. Furthermore, we introduce a hyperparameter $A \in \mathbb{R}_+$ which represents the amplitude. Hence, we get

$$\lambda_n = Af(\theta_n). \quad (7)$$

The amplitude A offers flexibility of avoiding excessively large regularization factors which could deteriorate the training procedure. Combining Equation (4) and (7) gives

$$L_\lambda(P) = L(P) + \sum_{n=1}^N (Af(\theta_n) \|\mathbf{w}_n\|_2^2). \quad (8)$$

¹We define $\|\mathbf{w}\|_2^2$ to denote the sum of squares of all elements also when \mathbf{w} is a matrix or a 3rd-order tensor.

TABLE I
COMPARISON OF THREE PERSON RE-IDENTIFICATION DATASETS, NAMELY, MARKET-1501 [1], DUKEMTMC-REID [2] AND MSMT17 [3].

Dataset	Market-1501	DukeMTMC-reID	MSMT17
Train Samples	12,936	16,522	32,621
Train Identities	751	702	1,041
Test Query Samples	3,368	2,228	11,659
Test Gallery Samples	15,913	17,661	82,161
Test Identities	751	1,110	3,060
Cameras	6	8	15

IV. EXPERIMENTS

In this section, we explain datasets, evaluation metrics and comprehensive analysis of our proposed method.

A. Datasets

We conduct experiments on three person re-identification datasets, namely, Market-1501 [1], DukeMTMC-reID [2] and MSMT17 [3]. Table I makes a comparison of those datasets. The MSMT17 dataset outshines the other two due to its large scale.

The Market-1501 dataset is collected with six different cameras in total. It contains 32,217 images from 1,501 pedestrians, and at least two cameras capture each pedestrian. The training set includes 751 pedestrians with 12,936 images, while the test set consists of the remaining images from 750 pedestrians and one distractor class.

The DukeMTMC-reID dataset includes 1,404 pedestrians that appear in at least two cameras and 408 pedestrians that appear only in one camera. The training and test sets contain 16,522 and 19,889 images, respectively. The query and gallery samples in the test set are randomly split.

The MSMT17 dataset is the largest person re-identification dataset which is publicly available, as of July 2020. It contains 126,441 images from 4,101 pedestrians, while 3 indoor cameras and 12 outdoor cameras are employed. In particular, the test set has approximately three times as much samples as the training set. Such setting motivates the research community to leverage a limited number of training samples that are available since data annotation is costly.

B. Evaluation metrics

Following the practices in [1], two evaluation metrics are applied to measure the performance, *i.e.*, mean Average Precision (mAP), and Cumulative Matching Characteristic (CMC) rank-k accuracy. The metrics take the distance matrix between query and gallery samples, in conjunction with the ground truth identities and camera IDs as input arguments. Gallery samples are discarded if they have been taken from the same camera as the query sample. As a result, greater emphasis is laid on the performance in the cross-camera setting.

Since the query samples may have multiple ground truth matches in the gallery set, mAP is preferable than rank-k accuracy for the reason that mAP considers both precision and recall.

TABLE II
PERFORMANCE COMPARISONS AMONG BASELINE, ADAPTIVEREID AND EXISTING APPROACHES. THE MAP SCORE ON MSMT17 IS THE MOST RELIABLE INDICATOR OF PERFORMANCE. R1: RANK-1 ACCURACY. -: NOT AVAILABLE. \dagger : RE-RANKING [10] IS APPLIED.

Method	Venue	Backbone	Market-1501		DukeMTMC-reID		MSMT17	
			mAP	R1	mAP	R1	mAP	R1
Annotators [27]	arXiv 2017	-	-	93.5	-	-	-	-
PCB [5]	ECCV 2018	ResNet50	81.6	93.8	69.2	83.3	-	-
IANet [28]	CVPR 2019	ResNet50	83.1	94.4	73.4	87.1	46.8	75.5
AANet [29]	CVPR 2019	ResNet50	82.5	93.9	72.6	86.4	-	-
CAMA [30]	CVPR 2019	ResNet50	84.5	94.7	72.9	85.8	-	-
DGNet [31]	CVPR 2019	ResNet50	86.0	94.8	74.8	86.6	52.3	77.2
OSNet [32]	ICCV 2019	OSNet	84.9	94.8	73.5	88.6	52.9	78.7
MHN [33]	ICCV 2019	ResNet50	85.0	95.1	77.2	89.1	-	-
BDB [9]	ICCV 2019	ResNet50	86.7	95.3	76.0	89.0	-	-
BAT-net [34]	ICCV 2019	GoogLeNet	87.4	95.1	77.3	87.7	56.8	79.5
SNR [35]	CVPR 2020	ResNet50	84.7	94.4	72.9	84.4	-	-
HOReID [36]	CVPR 2020	ResNet50	84.9	94.2	75.6	86.9	-	-
RGA-SC [37]	CVPR 2020	ResNet50	88.4	96.1	-	-	57.5	80.3
SCSN [38]	CVPR 2020	ResNet50	88.5	95.7	79.0	91.0	58.5	83.8
Baseline (Ours)	-	ResNet50	87.2	94.6	78.9	88.0	57.7	79.1
AdaptiveReID (Ours)	-	ResNet50	88.3	95.3	79.9	88.9	59.4	79.6
		ResNet101	88.6	94.8	80.6	89.2	61.9	81.3
		ResNet152	88.9	95.6	81.0	90.2	62.2	81.7
		ResNet152 \dagger	94.4	96.0	90.7	92.2	76.7	84.9

TABLE III

ABLATION STUDY OF BASELINE USING THE RESNET50 BACKBONE.
R1: RANK-1 ACCURACY.

Method	Market-1501		DukeMTMC-reID		MSMT17	
	mAP	R1	mAP	R1	mAP	R1
Minimal Setup	28.3	60.0	28.7	49.9	11.4	34.0
+ Triplet Loss	79.9	92.0	68.8	82.2	44.0	70.9
+ Regional Branches	81.3	93.3	71.2	84.2	47.9	74.2
+ Random Erasing	85.8	94.4	76.6	87.0	54.1	77.0
+ Clipping	86.8	94.3	78.1	87.6	56.5	78.4
+ L_2 regularization	87.2	94.6	78.9	88.0	57.7	79.1

C. Ablation study of baseline

The baseline differs from the minimal setup in five aspects, as discussed in Section III-B. Table III presents an ablation study to demonstrate how each component contributes to the performance on person re-identification. On the one hand, the triplet loss [22] brings the most significant improvements on all three datasets. The boost is due to the fact that the triplet loss is applied to the feature embeddings which are retrieved in the inference procedure (see Figure 1). Since the triplet loss directly optimizes the model in a manner comparable to similarity search, it closes the gap between the training and inference procedures. On the other hand, the other four components bring moderate improvements. It is conceivable that the model reaches better generalization by using hand-picked L_2 regularization factors which remain constant throughout the training procedure.

D. Comparisons with existing approaches

Table II shows performance comparisons among baseline, AdaptiveReID and existing approaches.

Firstly, all methods listed in Table II have surpassed the best-performing human annotators [27] on the Market-1501 dataset. In light of the scale of the Market-1501 and DukeMTMC-reID datasets (see Table I), these two small-scale datasets might have been saturated and more emphasis should be put on the MSMT17 dataset. Since mAP is preferable than rank-k accuracy, the mAP score on MSMT17 is the most reliable indicator of performance.

Secondly, our AdaptiveReID models are trained with the proposed adaptive L_2 regularization mechanism. The amplitude A in Equation 8 is set to 0.005. On the one hand, the AdaptiveReID method achieves decent improvements over the baseline method, especially on MSMT17 in which the mAP score increases by 1.7%. On the other hand, the AdaptiveReID method obtains the state-of-the-art performance on DukeMTMC-reID and MSMT17, very close to state-of-the-art performance on Market-1501, among methods which utilize the ResNet50 backbone.

Last but not least, deeper backbones (*i.e.*, ResNet101 and ResNet152) further improve the performance, at the cost of extra computations. Attributed to the re-ranking [10] method which exploits the test data in the inference procedure, new milestones have been accomplished, *i.e.*, the mAP scores on Market-1501, DukeMTMC-reID and MSMT17 stand at 94.4%, 90.7% and 76.7%, respectively.

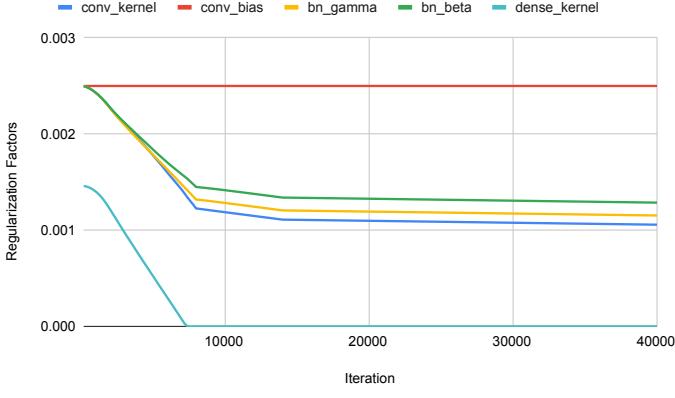


Fig. 3. The median value of regularization factors in each category, with respect to the number of iterations.

E. Quantitative analysis of regularization factors

Depending on the associated distinct parameter (see Equation 2), the regularization factors can be classified into five categories: *conv_kernel*, *conv_bias*, *bn_gamma*, *bn_beta* and *dense_kernel*, where *conv*, *bn* and *dense* denote the convolutional, batch normalization and fully-connected layers. In the following, we examine the regularization factors for a model trained on MSMT17 using the ResNet50 backbone.

Figure 3 visualizes the median value of regularization factors in each category, with respect to the number of iterations. Note that the learning rate gets reduced at iterations 8000 and 14000. While *conv_kernel*, *bn_gamma* and *bn_beta* behave similarly, *conv_bias* remains constant throughout the training procedure and *dense_kernel* drops to 0 in the early stage.

Figure 4 demonstrates a histogram of regularization factors in the last epoch, *i.e.*, the training procedure completes. The interval [0, 0.0025] is divided evenly into five buckets. For regularization factors from the same category, the values could differ significantly, *e.g.*, 2 and 38 regularization factors from *conv_bias* fall within the interval [0, 0.0005] and [0.0020, 0.0025], respectively. To be specific, the regularization factors from *conv_bias* in the two *Reduction* blocks are 0 (see Figure 2). If omitting the effects of the *Clipping* layer in Figure 1, those convolutional layers are followed by batch normalization layers which intrinsically cancels out the bias terms in aforementioned convolutional layers. Consequently, such regularization factors would converge to 0. In summary, this phenomenon reflects the superiority of our proposed method, in which each regularization factor is optimized separately.

F. Qualitative analysis of predictions

Figure 5 illustrates selected query samples with corresponding top 5 matches from the gallery set. Although query samples and erroneous matches may have similar appearances, minor differences could be observed under careful inspection, *e.g.*, the dissimilarity between backpacks. Furthermore, our models could retrieve correct matches even in the presence of large illumination changes, *e.g.*, the two examples from the MSMT17 dataset.

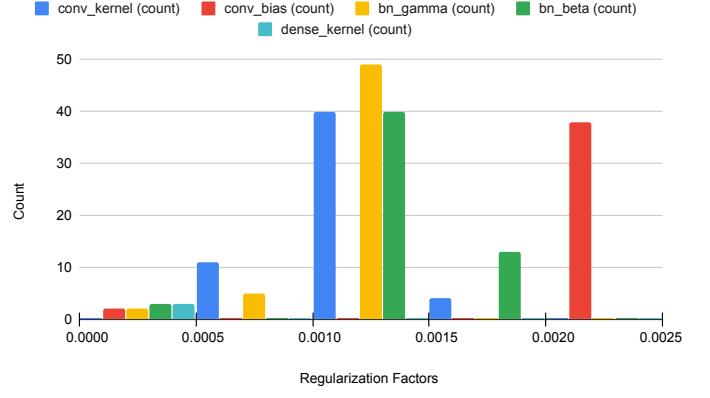


Fig. 4. Histogram of regularization factors in the last epoch.



Fig. 5. Selected query samples with corresponding top 5 matches from the gallery set. Images with orange, green and red border are query samples, correct matches and erroneous matches, respectively.

V. CONCLUSION

In this work, we revisit the L_2 regularization in neural networks and propose an adaptive mechanism named AdaptiveReID. Differentiated from existing approaches which employ hand-picked regularization factors that are constant, our proposed method can optimize those regularization factors adaptively through backpropagation. More specifically, we apply a scaled hard sigmoid function to trainable scalar variables, and use those as the regularization factors. Extensive experiments validate the effectiveness of our framework, and we obtain state-of-the-art performance on MSMT17 which is the largest person re-identification dataset.

REFERENCES

- [1] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.
- [2] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in *European Conference on Computer Vision*. Springer, 2016, pp. 17–35.
- [3] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer gan to bridge domain gap for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88.
- [4] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, “A siamese long short-term memory architecture for human re-identification,” in *European conference on computer vision*. Springer, 2016, pp. 135–153.
- [5] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 480–496.
- [6] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, “Deep attributes driven multi-camera person re-identification,” in *European conference on computer vision*. Springer, 2016, pp. 475–491.
- [7] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, “Improving person re-identification by attribute and identity learning,” *Pattern Recognition*, 2019.
- [8] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” *arXiv preprint arXiv:1708.04896*, 2017.
- [9] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, “Batch DropBlock network for person re-identification and beyond,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3691–3701.
- [10] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1318–1327.
- [11] J. Zhou, P. Yu, W. Tang, and Y. Wu, “Efficient online local metric adaptation via negative samples for person re-identification,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2420–2428.
- [12] K. Liu, B. Ma, W. Zhang, and R. Huang, “A spatio-temporal appearance representation for video-based pedestrian re-identification,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3810–3818.
- [13] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, “Video person re-identification by temporal residual learning,” *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1366–1377, 2018.
- [14] T. Van Laarhoven, “L₂ regularization versus batch and weight normalization,” *arXiv preprint arXiv:1706.05350*, 2017.
- [15] E. Hoffer, R. Banner, I. Golan, and D. Soudry, “Norm matters: efficient and accurate normalization schemes in deep networks,” in *Advances in Neural Information Processing Systems*, 2018, pp. 2160–2170.
- [16] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2018.
- [17] A. Lewkowycz and G. Gur-Ari, “On the training dynamics of deep networks with L₂ regularization,” *arXiv preprint arXiv:2006.08643*, 2020.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [19] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167*, 2015.
- [20] T. Salimans and D. P. Kingma, “Weight normalization: A simple reparameterization to accelerate training of deep neural networks,” in *Advances in neural information processing systems*, 2016, pp. 901–909.
- [21] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [22] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” *arXiv preprint arXiv:1703.07737*, 2017.
- [23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.
- [24] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” *arXiv preprint arXiv:1908.03265*, 2019.
- [25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [26] A. Krizhevsky and G. Hinton, “Convolutional deep belief networks on cifar-10,” *Unpublished manuscript*, vol. 40, no. 7, pp. 1–9, 2010.
- [27] 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.
- [28] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, “Interaction-and-aggregation network for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9317–9326.
- [29] C.-P. Tay, S. Roy, and K.-H. Yap, “Aanet: Attribute attention network for person re-identifications,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7134–7143.
- [30] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, “Towards rich feature discovery with class activation maps augmentation for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1389–1398.
- [31] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, “Joint discriminative and generative learning for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2138–2147.
- [32] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, “Omni-scale feature learning for person re-identification,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3702–3712.
- [33] B. Chen, W. Deng, and J. Hu, “Mixed high-order attention network for person re-identification,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 371–381.
- [34] P. Fang, J. Zhou, S. K. Roy, L. Petersson, and M. Harandi, “Bilinear attention networks for person retrieval,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8030–8039.
- [35] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, “Style normalization and restitutio for generalizable person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3143–3152.
- [36] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, “High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6449–6458.
- [37] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, “Relation-Aware Global Attention for Person Re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3186–3195.
- [38] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, “Salience-Guided Cascaded Suppression Network for Person Re-Identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3300–3310.