International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Siamese Network: Person Re-Identification Using Deep Learning with PyTorch

Ms. Gorle Dhana Lakshmi[1], Mr. Mudidana Jeevaratnam[2], Mr. Mudidana Sukumar[3]

[1]Assistant Professor(C), Department of CSE, UCEK(A), JNTUK Kakinada
[2,3]Student, Department of CSE, UCEK(A), JNTUK Kakinada

Abstract

Learning is made challenging by problems with person re-identification (Re-ID), like occlusion and misalignment. To rectify these problems, it's very essential to emphasize strong characteristics in intra-class variance. The attention-based Re-ID techniques used today ignore unique characteristics in favour of commonalities. In this research presents a novel Siamese network for person's Re-ID on attentive learning. In contrast to prior techniques, the leveraged the Siamese network's characteristics to create the attention-loss and attention module that focus attention on recognizable and common elements. The encoder-decoder attention module is to determine the general form of the body and channel attention to choose significant channels. The term "uniformity loss" now describes the attention deficit brought on by the triplet loss. A distinct attention map that emphasizes both common and discriminative characteristics is produced because of the homogeneity loss. The recommended network performs better than the most recent techniques on the Market-1501, according to a comprehensive testing procedure.

Keywords: Re-identification, Siamese network, Market-1501 data sets.

I. INTRODUCTION

One kind of neural network architecture utilised for tasks involving determining the similarities or differences between two input samples is the Siamese Network. The network is made up of two identical subnetworks that are trained concurrently and possess the right sets.

The Siamese network's fundamental goal is to determine a similarity metric between two input samples. Put another way, the network is built to return a low value when the two input samples are not comparable and a high value when they are. As a result, it is used for text or image similarity matching, face recognition, and signature verification. When it comes to computer vision, re-identification (Re-ID) of persons is an important field of study. Person Re-ID in a surveillance system with many cameras aims to find a specific individual from multiple non-overlapping camera perspectives. Person Re-ID has therefore been extensively used in Applications for video analytics, like multi-target tracking [1] and human recovery [2]. Siamese networks are beneficial in situations where data is limited since they require extremely few instances of training data. This is their main advantage. Furthermore, because the two subnetworks share heights, the model can adapt to new inputs with greater success.

Traditional person Re-ID tasks often utilise hand-crafted features [5] and distance metric learning [3], focusing on edge histograms and colour features [7]. These methods extract HSV colour histograms and

emphasise body shape edges through image segmentation.

Siamese networks have proven to perform effectively in several different contexts, such as speech recognition, image identification, and classification. They have also been used for natural language processing tasks, including paraphrasing and recognising phrase similarities.

The Re-ID exceeded the use of benchmarks (12), and deep learning has progressed. Global feature representation (GSP) learning was used by early deep re-ID algorithms to integrate picture categorisation into the re-ID issue (13). Larger intra-class variances, occlusion, background clutter, and other fundamental problems that Re-ID faces are not addressed by the basic structure of GSP-based classification networks. By concentrating on the deterministic region, attention strategies inevitably incorporate the trait that they deem significant. The person in the first and second rows is the same, as Figure (1a) shows. On the other hand, the bag is missing from the second row, where it wouldn't be as noticeable. However, as Figure 1b illustrates, the bag disappeared at a new camera angle, causing the shirts and shoes to disappear as well, indicating that the bag would have lost any distinctive qualities as well. However, shoes and shirts were the focal points of the suggested activation maps for (a) and (b), respectively. In this particular camera perspective, the bags are considered noteworthy, but not in others. In light of Re-ID's qualities, uniformity loss was therefore required in order to concentrate on common and discriminative aspects. Uniformity loss, which minimises variations in attention mappings between items, has been used to achieve this. The following is a summary of this work's primary contributions:

- Currently available is a focused, Siamese network for learning individuals' Re-ID built on a shared discriminator architecture, a learning-based Siamese network. Our technique consists of two modules: a channel attention module and a robust feature extraction encoder-decoder attention module.
- Uniformity loss is developed to learn features that aid in the Siamese network's accurate learning of more significant features, both discriminative and common.
- Our proposition is supported by extensive studies conducted on the three most commonly used benchmarks, using equivalent evaluation techniques that are both objective and subjective.

Figure-1: Activation maps using the Market1501 dataset for visualization.

II. RELATED WORKS

A. Mechanism

Through the use of a neural network, targeted regions are identified, and spatial pinpointing is carried out via the attention process. The first inspiration for this attention mechanism came from studies conducted on natural language processing. But during training, the recurrent attention model encounters

what is known as the "hard attention problem," which is caused by its inability to focus on a single place in the image. Bahdanau and colleagues developed a "soft attention model" that investigates all input aspects on it. With this method, the RNN limitations encoder-decoder network's fixed-length vector encapsulating all sentence information are removed. In the field of natural language processing, translation performance was greatly improved by Bahdanau's approach.

Recent developments in attention mechanisms have had a big impact on a lot of different computer vision applications. An attention-based image captioning model was presented by Xu and colleagues, which allows the machine to learn to describe a picture on its own. An attention model for image categorization was suggested by Sermanet and colleagues. The model learns to recognize high-resolution attention areas and extracts important parts from the image to help in discrimination. Li and associates also expanded the use of attention models to the domain of object detection. Their approach uses an attention model to take into account both local and global settings. Using both internal and exterior local contextual information,

In order to maximize the alignment of mismatched images, Li and colleagues created the Harmonious Attention CNN (HA-CNN), a model that simultaneously learn to pay attention to soft and hard pixels. Li et al. also presented a spatiotemporal attention model designed to recognize different body parts. However, it is important to remember that these techniques.

B. Person ReID Siamese Network

An NN architecture that can recognize in-person re-identification (ReID) situations is called a Siamese net. Determining the degree of similarity or difference between two pairs of input data is the goal of the Siamese network. compare if two photos, usually of the same person taken from multiple camera angles, reflect the same individual or not is the aim of person ReID.

The attention loss mechanism does two things. Firstly, it makes it easier to remember consistent features, which makes it possible to recognize shared characteristics between people. Second, it helps differentiate certain features from others, which improves the network's ability to detect finer details. Furthermore, enhanced spatial localization is directly facilitated by the intrinsic structure of the Siamese network, which facilitates a smooth learning process from beginning to end.

The architectural design not only the advantages most advanced the attention models, but also leverages their capacity to extract the features which are specifically well-suited by the person re-identification (Re-ID) tasks.

III. THE PROPOSED METHOD

They provide a novel attention-based learning network in this part that is intended for person re-identification (Re-ID) tasks. Our technique combines identification, verification, and uniformity—three crucial loss components—using a Siamese network design.

Every branch of the network, shown as in Figure 2, produces final the feature representation, represented by "f," which functions feature extractor and then employed in the attention methods. To forecast values for each loss component, additional processing is applied to the feature representation "f". Further layers, such as fully-connected layers, global average pooling, and spatial attention mechanisms, are involved in this prediction.

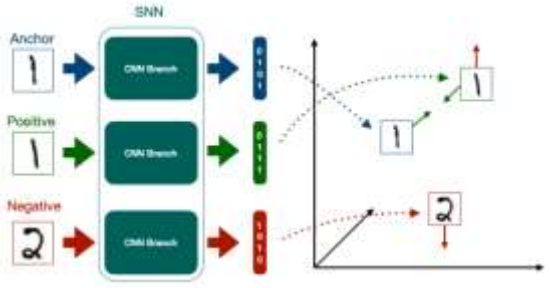


Figure 2: The architecture of the suggested techniques.

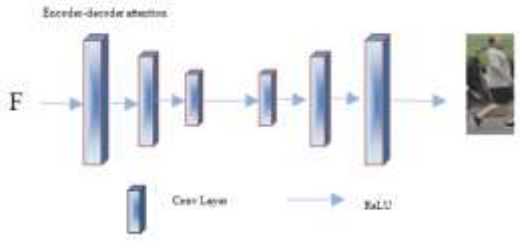
A. Identifying the Attention Module

When combined, these components create the discriminant features and incorporate attention strategies to draw attention spatial to pertinent information. Concurrently, the identification module extracts attention regions by using feature maps that come from the convolutional layer.

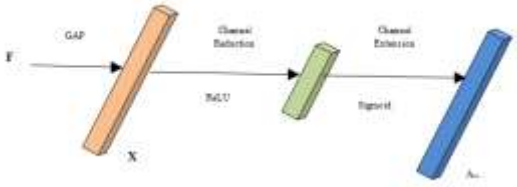
1. Attention Module of Encoder-Decoder

The architecture of the encoder- decoder provides pixel-wise predictions and information capture at many scales. In order to extract semantic information, the encoder first others the input resolution; the decoder then up samples the data to bring it back to its original size. Through the combination of this structure's input and output, the simultaneously acquire information at many scales.

Three layers of convolution are followed by three layers of deconvolution in the encoder and decoder attention branch. In order to create a gentle attention mask, this attention module is essential in turning off background noise and concentrating on the general contour of the body. The path of the encoder revised the unpadded of the convolution, followed by ReLU activation, 3 times, as shown in Figure 3. The last deconvolution layer in the decoder circuit reduces the channel dimension to one while repeating deconvolution and ReLU activation three times.



(a)



(b)

Figure 3: a) Encoder-decoder b) attention modules channel

2. Attention Module Channel

Stop paying attention to spatial information and instead concentrate on comprehending the inter-channel

relationships inside the feature map within the channel attention branch. To do this, a technique known as pooling is used to create a feature vector, represented by the letter "x," thereby flattening the feature map "F." Equation (1) shows that the content of "x" is obtained by dividing each channel in feature map "F" by the channel size (c).

The dimension reduction layer are the main parts of the channel attention operator, or "fch." The channel attention feature vector is finally produced by the sigmoid activation function, which is essential for allocating importance scores to each channel Ach∈Rc×1×1.

W1∈$\mathbb{R}\frac{c}{r}$×C and W2∈$\mathbb{R}$C×$\frac{c}{r}$ are representations of expansion layers' decrease with a reduce the ratio of r. The ideal reduction ratio in experiments was suggested to be r=16. To get the final feature map F^ encoder-decoder attention vector.

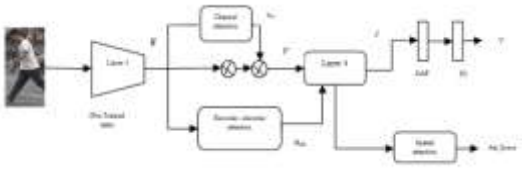


Figure 4: Proposed System Pipeline

The following is an illustration of our suggested method's process flow:

- **F**: This is the feature map that was produced following the ResNet50 model's conv3 layer processing.
- **Ac** This is the feature map that was produced following the ResNet50 model's conv3 layer processing.
- **Aed**: The spatial attention map derived from the encoder-decoder attention module.
- **zi**: A feature vector for identity (ID) verification and prediction.
- **Attention Score**: Using a Siamese network, this score is calculated by comparing the attention maps (Ac and Aed)

This pipeline demonstrates how our approach combines elements of Siamese networks with attention processes to handle human re-identification tasks, such as identity verification and prediction.

3. Identification Loss

The suggested feature extractor, layer 4, is positioned behind the identification classifier. The identification predictions produced by this classifier—which consists of two fully-connected layers—are predicated on the input picture feature vector "f." Put more simply, "f" uses the identification classifier to "y^".

One of the objectives of this optimization process is to reduce the identification loss, or "Lid." In formal terms, the variables "y^" and "Lid" have the following definitions:

$\hat{y}_i = \frac{WI\ o\ fi}{\sum_j WI\ o\ fi}$

where wI stands for the identification classifier's layers' parameters. The label is represented by yi, and its prediction is Þi.

IV. EXPERIMENTS

A. Evaluation Metrics & Datasets

Collection to evaluate our methods using three extensive person re-identification (Re-ID) benchmark datasets:

• **Market1501 [12]:** includes 32,668 individual photos total, representing 1501 distinct identities, taken from 6 different camera angles. There are 12,936 picture representations for 751 individuals in the training set

These datasets provide thorough benchmarks for evaluating our method's performance and efficacy in the person re-identification domain.

Evaluation Metrics. We use two key performance indicators in our assessment, which the apply to all datasets: the mAP and the CMC. These metrics measure following:

- **Cumulative Matching Characteristic (CMC):** By calculating the actual matches inside the top n-ranking, the CMC curve measures the accuracy of matching.
- **Mean Average Precision (mAP):** Precision over the whole range of recovered results is taken into account when calculating the mAP measure, which assesses the method's overall effectiveness. It considers the quality of matches over the whole list of retrieved results in addition to the accuracy of the top-ranked matches.

When taken as a whole, these measures offer a strong assessment of the method's effectiveness in person re-identification, capturing both ranking accuracy and match quality across different datasets and contexts.

B. Implementation Details

As the foundation layer of our methodology, the employ the Market1501 pre-trained model, which was developed on the original ImageNet dataset. They added the attention module to Market1501's third residual block in order to integrate it into the architecture.

The main actions and specifics of our model are as follows:

- **Train and import data**: In the training phase, the feature vector "z" was employed for image pair comparison in addition to label prediction using a fully-connected layer. Every training batch included two positive and negative samples chosen at random for each of the P identities.
- **Configuration**: We used the embedding network to extract the feature vector for every image pair in order to verify our findings. Next, these vectors there contrasted with each P identity's feature vectors.
- **Similarity Optimization**: To maximize the uniformity, it additionally computes attention scores for P identities and every image pair using the L1-norm.
- **Create APN Dataset**: For consistency, all human pictures were shrunk to 256 x 128 pixels.
- **Load datasets and Crete Model**: For training, the employed the SGD algorithm for a starting learning rate with 0.04. They employed a medium momentum with 0.9 and found that learning rate declined by 0.1 years across 30 epochs.
- **Create Train and Eval Functions**: Establish the train function models for the criterion, optimizer, and data loader. Distinguish between anchor devices and the positive and negative rates.
- **Create Training Loop**: The data on train loss and test loss is obtained using a model, train loader, optimizer, and criteria with a valid loss. If the condition is smaller than the rate, the best model is saved.
- **Test Phase**: In the network, processed only P IDs of the test batch without taking image pairs into account, and it solely engaged the embedding network.

The model's efficiency and effectiveness in person re-identification there demonstrated by the consistent application of these parameters and configurations across all three datasets.

C. A Comparison of Network Architectural Change with Anchor Embeddings:

The attention module was implemented at various backbone network scales in an experiment to examine the impact of the suggested uniform attentive learning-based Siamese network. The attention module's application in layer 3 produced the best results. To use the effective attention module, one must select a feature map that appropriately balances coarse and fine information. The complete body shape cannot be seen by the encoder-decoder focus due to the small receptive fields of layer 1 and layer 2. For learning discriminative features, the encoder-decoder attention map is unsuitable since layer 4 has a relatively small feature map. The input is received by the encoder of the encoder-decoder attention.

D. Qualitative Inferential Analysis

The activation map was validated qualitatively, and the method you recommended was reassessed. The results of Market 1501 are shown in 2nd column of Figure 5, and the employment of the third column displays only the attention module without compromising consistency. The fourth column displays the result of the recommended process.

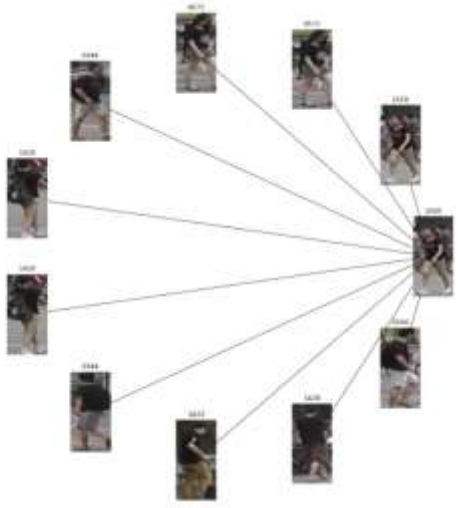


Figure 5: Visualized Examples: Using the market1501 dataset, compare the recommended method with the others: Enter image data and only use the focus module. Recommended Method

The scenario where the invisible items become apparent when the camera angle changes is depicted in Figure 5. In this case, having possessions is not a normal characteristic. The bag was recognized in the second column as an essential part of the third row even though it wasn't visible the views of front. Items in the 3rd and 4th columns caught our attention in addition to the bag. The occlusion scenario with obstructions is depicted in Figure 5. The 2nd column in each pair is the image's centre, ignoring any obstacles.

V. CONCLUSIONS

To extract deterministic characteristics, this paper illustrates a Siamese network based on ensemble attentive learning. The Siamese network's channel attention module and encoder-decoder attention module make up the suggested network. The attention module encoder-decoder does not see areas of the body; it sees the entire body. Additionally, the suggested network suggests the uniformity loss by

leveraging the characteristics of these attention modules. This makes it possible to concentrate on more predictable zones and to be resistant to posture variation and occlusion issues. Numerous trials show that the suggested network performs better than the most advanced methods for human re-identification, both quantitatively and qualitatively.

References

1. Wang X. Intelligent multi-camera video surveillance: A review. *Pattern Recognit. Lett.* 2013. doi: 10.1016. [Google Scholar]
2. Loy C.C., Xiang T., Multi-camera activity correlation analysis; Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; Miami, USA. 20–25 June 2009; pp. 1988–1995. [Google Scholar]
3. Theinberger K.Q., Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* 2009 [Google Scholar]
4. Zheng W.S., Gong S., Reidentification by relative distance comparison. *IEEE Trans. Pattern Anal. Mach. Intell.* 2012;doi: 10.1109/TPAMI.2012.138. [PubMed] [CrossRef] [Google Scholar]
5. Zajdel W., Keeping track of humans: Have I seen this person before? IEEE International Conference on Robotics Automation; Barcelona, Spain. 18–22 April 2005; pp. 2081–2086. [Google Scholar]
6. Gray D., Tao H. *European Conference Computer Vision.* Springer; Viewpoint invariant pedestrian recognition with an ensemble of localized features; pp. 262–275. [Google Scholar]
7. Farenzena M., Bazzani L., Perina A., Murino V. Person re-identification by symmetry-driven accumulation of local features; IEEE Computer Society Conference on Computer Vision and Pattern Recognition; San Francisco, CA, USA. 13–18 June 2010; pp. 2360–2367. [Google Scholar]
8. Gheissari N., Sebastian T.B., Person reidentification using spatiotemporal appearance; IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) USA. 17–22 June 2006; pp. 1528–1535. [Google Scholar]
9. Zhao R., Ouyang W., Unsupervised salience learning for person re-identification; IEEE Conference on Computer Vision and Pattern Recognition; Portland, OR, USA. 23–28 June 2013; pp. 3586–3593. [Google Scholar]
10. Mignon A., Jurie A new approach for distance learning from sparse pairwise constraints; IEEE Conference on Computer Vision and Pattern Recognition; Providence, RI, USA. 18–20 June 2012; pp. 2666–2672. [Google Scholar]
11. Liao S., Hu Y., Zhu X., Person re-identification by local maximal occurrence representation and metric learning; IEEE Conference on Computer Vision and Pattern Recognition; Boston, MA, USA. 7–12 June 2015; pp. 2197–2206. [Google Scholar]
12. Zheng L., Shen L., Tian L., Wang S., Wang J. Scalable person re-identification: A benchmark; IEEE International Conference on Computer Vision; Santiago, Chile. 11–18 December 2015; pp. 1116–1124. [Google Scholar]
13. Li W., Zhao R., Xiao T., Deepreid: Deep filter pairing neural network for person re-identification; IEEE conference on computer vision and pattern Recognition; Columbus, OH, USA. 24–27 June 2014; pp. 152–159. [Google Scholar]
14. Zheng Z., Zheng L., Unlabeled samples generated by gan improve the person re-identification baseline in vitro; IEEE International Conference on Computer Vision; Venice, Italy. 22–29 October 2017; pp.. [Google Scholar]

15. Krizhevsky A., SutskeverI.,.Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 2012; doi: 10.1145/3065386. [CrossRef] [Google Scholar]
16. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv.* 20141409.1556 [Google Scholar]
17. Szegedy C., Ioffe S., Vanhoucke V., Alemi A.A. Inception-v4, inception-resnet and the impact of residual connections on learning; AAAI Conference on Artificial Intelligence; San Francisco, CA, USA. 4–10 February 2017. [Google Scholar]
18. Huang G., Liu Z., Van Der Maaten L., Densely connected convolutional networks; IEEE Conference on Computer Vision and Pattern Recognition; Honolulu, HI, USA. 21–26 July 2017; pp. 4700–4708. [Google Scholar]
19. Hu J., Shen L., Squeeze-and-excitation networks; IEEE Conference on Computer Vision and Pattern Recognition; Salt Lake, UT, USA. 18–22 June 2018; pp. 7132–7141. [Google Scholar]
20. He K., Zhang X., Ren S., Deep residual learning for image recognition; IEEE Conference on Computer Vision and Pattern Recognition; Las Vegas, NV, USA. 26 June–1 July 2016; pp. 770–778. [Google Scholar]