Science Arena Publications
Specialty Journal of Electronic and Computer Sciences
ISSN: 2412-7485
Available online at www.sciarena.com
2018, Vol, 4 (2): 14-20

Developing a Model of Deep Learning Architecture Using Generalized and Responsive Pooling for Person Re_identification

Hourieh Sadat Jamalidinan*, Mahmood Fathy, Ahmad Akbari

Faculty of Computer Engineering, Iran University of Science and Technology, Narmak, Tehran 16846-13114, Iran.

***Corresponding Author**

Abstract: *Nowadays, deep learning networks have drawn a lot of attentions in the area of person re-identification. It is attempted in the present article to improve a deep learning neural network so that it can gain more distinct and better characteristics and reach a higher accuracy level. The network simultaneously deals with feature learning and subsequent feature comparison. The use of generalized pooling layer in lieu of ordinary pooling layer has been an innovation proposed in this article. There are two types of generalized pooling used in this network. Tree pooling has been the method of choice in the initial layers and the proposed network is trained with pooling filters and a combination of the generalized pooling and pooling filters so that the network could be responsive. In the final layers, we have used gated max-average pooling with which the network is trained via a gating mask during the learning process so as to finally reach a relative composition of the two types of pooling, i.e. max pooling and average pooling. The network has succeeded in acquiring very much better results on such large datasets as CUHK03 and CUHK01 in comparison to its preliminary state; the network also offers higher accuracy on such smaller datasets as VIPeR in contrast to its mainstream method but the accuracy enhancement is rather trivial due to the few numbers of the data.*

Keywords: *Deep Learning, Convolution Neural Networks, Pooling, Machin Vision, Person Re-identification*

INTRODUCTION

Machine learning investigates methods enabling computers to be trained with data for presenting better performance. The goal in machine learning is that a computer can reach a higher level of output gradually with increasing the number of the data.

Deep learning has been focused on by a great number of researchers during the recent years. Deep learning is widely applied in various fields of machine learning. The majority of the studies carried out in this regard are based on deep learning neural networks. These studies incorporate network architectures definition, training methods explanation and networks efficiency evaluation. The latter is usually done via training the network with various training data followed by evaluation of the learnt data. There are also many tools developed for deep learning.

Discussions posited concerning deep neural networks are drawn on studies conducted on artificial neural networks. In fact, deep learning has offered an optimal performance in various areas of artificial intelligence and machine learning through the elevation of computers' calculation power and the possibility of creating and training massive networks using a large number of indicators.

In fact, deep learning refers to computational models comprised of several processing layers. These layers rate the data representation of a network via creating various learning abstraction levels. It can be generally stated that deep learning is applicable to various fields of machine learning such as sound and speech processing, language modeling and natural language processing, information retrieval, machine vision and multi-objective learning (LeCun et al., 2015).

We are in an era of enormous dispersion of data in the global internet networks. Nowadays, there is a huge volume of unstructured data in internet environment.

Research on deep learning is now an active research area. Some of these studies have dealt with machine vision. One fundamental issue in regard of machine vision is automatic processing of video and image. Inspired by hierarchical processing of data in brain as well as the way of humans neural network's storing of the information (Bengio, 2009), deep learning seeks better and faster actualization of processing in this area.

An important problem in the field of image and video processing is person re-identification that has been transformed into an important and applied issue in machine vision domain. In person re-identification, the system should be able to identify an individual by processing various images and videos recorded by various cameras. The individual's taking of various postures in different situations makes the recognition a very complex task.

Currently, there is an extensive network of cameras used in airports, train halls and university campuses and administrative buildings and others. Covering geographical areas, these cameras do not share a field of vision so that their monitoring of the covered area could be corroborated. These networks provide a large volume of video and image information that are manually used by security guards and/or the police and others for saying legal objectives. The observation and monitoring of these videos needs a lot of time and capital. Automatization of video and image processing can speed up processing and enhance its quality (Ahmed, Ejaz et al., 2015).

Generally, two essential actions should take place for person re-identification: 1) procuring of image features (Zha et al., 2014; Yang et al., 2014; Zhang et al., 2014; Bromley et al., 1993) attainment of similarity metrics for the comparison of features with one another (Koestinger et al., 2012; Li and Wang, 2013; Li et al., 2013; Martinel et al., 2014; Zhao et al., 2013). The present study deals with deep learning neural networks covering both of the aforesaid actions at the same time.

Convolution networks have been developed with numerous modules one of which is pooling layer which has major role in size reduction. In 2016, Chen-Yu Lee added several parameters to a network that smartened the pooling layer. These parameters were taught during network training process and provided for the improvement of network performance and enhancement of the accuracy levels. In this article, we have proposed a deep learning neural network. The technique used is based on inception methods and using smart pooling layer instead of ordinary and primary layers and uses fewer free parameters than those utilized in RESNET networks. These types of networks were selected because of having a lower number of free parameters that prevents the network from appearing overfit.

Related Works:

With increasing the use of deep learning neural networks in image classification led Yi et al. (2014) to the investigation of person-re-identification for the first time in 2014. They employed a Siamese Convolution Neural Network (Bromley et al., 1993) to conduct deep metrics learning. In this network, the input image is divided into three horizontal parts overlapping one another and each part passed through two convolution layers along with a fully connected layer. The network used cosine similarity scale. In the same year, Li et al. (2014) introduced another deep learning neural network in person re-identification area that was based on Siamese network. The network had been trained based on filter pairs for encoding photometric conversions. In 2015, Ahmed et al. (2015) offered another deep learning neural network that was an improved model of Siamese network. In this network, the feature difference of two adjacent images is obtained by mapping the feature differences of the first image on those of the neighborhood of the second image for a given spot. Tied convolution is used in the network for attaining higher level attributes of the image in such a way that the input image is

inserted into two convolution layers featuring max pooling and the created mapping is entered into the cross-input neighborhood difference layer after which the feature differences between two images can be extracted. Then, the layer output is sent to a ReLu layer. In the next layer, i.e. patch summary features, a summary of the higher level features is learnt via convolution operation the output of which is sent to ReLu layer. The spatial relations of the two input images are obtained from previous stage mapping in across-patch features layer. At the end, the output of the previous layer is sent to higher-order relationship layer in which the inputs are inserted into softmax function that gives the final output. During the recent years, Siamese network has been usually employed in identity recognition area for the limited nature of training data. Nowadays, the increase in the number of the various datasets in the human identification field has provided for the improvement of the training process but, even with the different datasets featuring various sizes and domains in the area of person re-identification, none of the approaches is good enough alone for training the network (Xiao et al., 2016). Due to the same reason, in 2016, Mr. Tong Xiao introduced a method for combining various datasets in the area of person re-identification and proposed a novel method of dropout. The enlargement of the volume of the network training data provided for a more generalized way of network training that reduces the network's underfit. In the current research paper, the network proposed by Mr. Ahmed et al has been modified with smart pooling instead of ordinary pooling and applied to person re-identification and accuracies were greatly enhanced.

Proposed Architecture:

The performance of pooling operations has been revised in respect to common pooling methods such as stochastic, average and max pooling and various studies have presented different performance levels for such an operation (Boureau et al., 2010; Zeiler and Fergus, 2013).

In deep learning networks, a sort of pooling operation is used in pooling layers that is termed max or average pooling and the possibility of making use of more than one pooling method which aims to improve of system performance (Scherer et al., 2010).

Chen-Yu Lee in (2016) introduced a special type of pooling. In this method, three various models of pooling layers were applied as substitutes to ordinary pooling operations. The first model was composed of mixed max-average pooling in which a free parameter was added in order to simultaneously perform max and average pooling. The inherent characteristic of the region subject to pooling operation is disregarded and if the method is carried out foreach pooling layer of the network, there was a need for training a parameter during the network instruction process. Compared with the various image conversion methods, this pooling operation performs better. The formula of the method is as shown below:

$$F_{mix} = \alpha F_{max}(x) + (1-\alpha)F_{ave}(x) \qquad (1)$$

Where, $F_{max}(x)$ is equal to the max pooling performance and $F_{ave}(x)$ is equal to average pooling performance.

In gated max-average pooling model, for a pooling operation of the dimension of n×n, about n×n free parameters had to be added for gating mask learning for every pooling layer. As it can be seen in the proposed formula, the pooling operation considers the inherent properties of the region in which pooling is to take place hence it is found smarter in this regard. The output of gating mask multiplication and the region subject to pooling is sent to a sigmoid function.

$$f_{gate}(x) = \sigma(\omega^T x) f_{max}(x) + \big(1-\sigma(\omega^T x)\big) f_{avg}(x) \qquad (2)$$

Where, W is the amount of gating mask learnt in the course of training process. The deep learning model presented in (Ahmed, Ejaz et al., 2015), in which the pooling layer is applied in two stages, has been used herein. Max pooling with a 2×2 dimension is used in across patch feature layer that obtains the spatial relations of two images via neighborhood differences. In this layer, there is made use of gated max-average pooling instead of max pooling.

The proposed gated max-average pooling consists of two fixed pooling operations (max and average pooling). Another way of pooling generalization is allowing the learning of the operations to be combined. These separate pooling layers are convolution layers. This way, pooling conducts its operation in separate in each channel. The isolation of each channel means that fewer parameters can be produced in respect to the convolution layers created.

The simplest version of the method never combines the pooling operations learnt rather it merely learns the pooling operation within the format of amounts given in pooling filters. In tree pooling method, pooling filters as well as a responsive combination of them is to be learnt during learning process.

Both aspects of learning occur within the format of a binary tree. Each leaf denotes a filter learnt during training process and the middle nodes of the tree are the products of two child nodes' combination and this is continued to the tree's root at which point the final result is extracted (Figure 1). Each of these combinations in the interior nodes is a process learnt during instruction process.

For a leaf-like node m, pooling filter is designated by V_m. Assuming m as an internal node, gated max is designated by W_m. In this case, pooling result is obtained for each arbitrary node as demonstrated below.

$$f_m(x) = \begin{cases} V_m^T x & if\ lead\ node \\ \sigma(\omega_m^T x) f_{m,left}(x) + \left(1 - \sigma(\omega_m^T x)\right) f_{m,right}(x) & if\ internal\ node \end{cases}$$

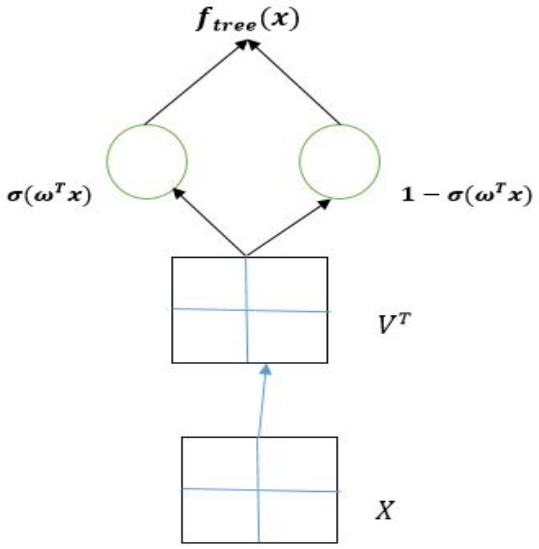

Figure 1: Tree Polling (2 levels), we indicate the region being pool by X, gating mask by W and pooling filter by V

After having convolution operation done, the output is inserted into a pooling operation in tied convolution layer. Max pooling was herein replaced by a two-level tree pooling.

Experiments:

Similar to primary networks, there is a pair of image used in this network the output of which determines whether they belong to an identity (positive) or identities (negative). Due to the reason that the number of positive and negative data largely differs and the data feature imbalance causing overfit, following the lead of the main study, we also made use of label-preserving conversion (Krizhevsky et al., 2012) to increase the number of the training data. Data augmentation and down-sampling problems of the negative data were also taken into account as was also done in the base article. The proposed model was implemented in CAFFE Framework (Jia et al., 2014) and three datasets were utilized for comparing the proposed method with its prior

version. The datasets used in this method are: CUHK01 (Li et al., 2012), CUHK03 (Li et al., 2013) and VIPeR (Gray et al., 2007). The obtained accuracy was evaluated by CMC scale in top-1 (Moon and Phillips, 2001). Similar to what has been done in the primary method, we also classified the data into three sets, namely test, train and validation. The latter two data were used in model instruction and the test data were applied for calculation of the accuracy of the model.

Classification Results:

CUHK03:
CUHK03 dataset incorporates 13164 images of 1360 passengers taken by six cameras. The dataset has been labeled both manually and automatically (Felzenszwalb et al., 2010). In this method, the accuracy was investigated on both dataset models. The method proposed in the present study reached an accuracy value equal to 67.4% while the maximum accuracy of the previous method for this dataset is 54.76%. In case of automatic labelling, the accuracy mean value was found increased from 44.96% to 51.3% in contrast to the mean value obtained for the primary method.
CUHK01:
CUHK01 dataset is composed of 971 identities. To make comparisons, similar to the prior work method, we also selected 100 identities once for testing and 871 identities for training. The mean accuracy was found increased from 65% to 68%. In a second state, about 486 identities were used for test and 485 identities for training. Overfit in training parameters added to the network occurred due to the small number of training data. To overcome the problem, the network was first trained on CUHK03 data following which fine-tuning was carried out on CUHK01 data. The accuracy was changed from a value equal to 47.5%, obtained in the base article, to 56%, obtained in the proposed method. As it can be seen, due to the low number of data and large number of free parameters in the network, the change in pooling layer did not provide for well-training of the network and the accuracy rates were not increased as good as the ones obtained for CUHK03.
VIPeR:
This dataset is comprised of 632 images taken in two views of the passersby. The dataset was divided into two halves, one for testing and the other for training. Overfitting problem came about again. Since the number of the data in this dataset is not very high, the network trained on CUHK03 and CUHK01 datasets was fine-tuned on VIPeR dataset and the accuracy was subsequently increased from 34.81% to 41.2%.

Conclusion:

We succeeded in proposing a method based on a model previously constructed by Ahmed et al based on features' neighborhood discrimination and used generalized pooling the performance of which depends on the inherent properties of the region subject to pooling to improve the method in such a manner that higher accuracies were obtained. The present study is mostly concentrated on the use of generalized pooling in deep learning neural networks pertaining to person re-identification. It was found out that smartening of the pooling in respect to the inherent attributes of the region subject thereto causes an increase in the number of free parameters in the network. To make up for this problem, there is a need for making use of a larger number of data for training because the use of such a type of pooling causes better and more distinct feature extraction following which the accuracies can be increased more.

References

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Associates, Inc., 2012. 6

2. Ahmed, Ejaz, Michael Jones, and Tim K. Marks. "An Improved Deep Learning Architecture for Person Re-identification." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): n. pag. Web
3. D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In in IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro, 2007. 6
4. D. Scherer, A.M¨uller, and S. Behnke. Evaluation of pooling operations in convolutional architectures for object ecognition. In ICANN. 2010.
5. D. Yi, Z. Lei, S. Liao, S. Z. Li et al., "Deep metric learning for person re-identification." In ICPR, vol. 2014, 2014, pp. 34–39.
6. H. Moon and P. J. Phillips. Computational and performance aspects of pca-based face-recognition algorithms. Percep-tion, 2001. 6
7. J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. S¨ackinger, and R. Shah, "Signature verification using a Siamese time delay neural network," International Journal of Pattern Recognition and Artificial Intelligence, vol. 7, no. 04, pp. 669–688, 1993.
8. Lee, Chen-Yu, Patrick W. Gallagher, and Zhuowen Tu. "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree." Artificial Intelligence and Statistics. 2016.
9. M. D. Zeiler and R. Fergus. Stochastic Pooling for Regularization of Deep Convolutional Neural Networks. arXiv preprint rXiv:1301.3557, 2013.
10. M. Koestinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In CVPR, 2012. 2, 6M. Koestinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In CVPR, 2012. 2, 6
11. N. Martinel, C. Micheloni, and G. Feresti. Saliency weighted features for person re-identification. In ECCV Workshop on Visual Surveillance and Re-identification, 2014. 2
12. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained partbased models. IEEE Trans. Pattern Anal. Mach. Intell., 32(9):1627–1645, Sept. 2010. 6
13. R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In CVPR, 2014. 2, 7, 8
14. R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In ICCV, 2013. 2, 7
15. W. Li and X. Wang. Locally aligned feature transforms across views. In CVPR, 2013. 2
16. W. Li, R. Zhao, and X. Wang. Human re-identification with transferred metric learning. In ACCV, 2012. 2, 6, 7
17. W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2014, pp. 152–159.
18. Xiao, Tong, et al. "Learning deep feature representations with domain guided dropout for person re-identification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
19. Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015.
20. Y. Bengio, "Learning Deep Architectures for AI," Foundations and Trends® in Machine Learning, vol. 2, no. 1, pp. 1-127, 2009.
21. Y. Boureau, J. Ponce, and Y. LeCun. A Theoretical Analysis of Feature Pooling in Visual Recognition. In ICML, 2010.
22. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014. 6
23. Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Li. Salient color names for person re-identification. In ECCV, 2014. 2

24. Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, and J. Smith. Learning locally-adaptive decision functions for person verification. In CVPR, 2013. 2, 8
25. Z. Zhang, Y. Chen, and V. Saligrama. A novel visual word co-occurrence model for person re-identification. In ECCV Workshop on Visual Surveillance and Re-identification, 2014. 2, 7