Does Pooling Really Matter? An Evaluation on Gait Recognition

Claudio Filipi Goncalves dos Santos[1](✉), Thierry Pinheiro Moreira[2],
Danilo Colombo[3](✉), and João Paulo Papa[2](✉)

[1] Federal University of São Carlos - UFSCar, São Carlos, Brazil
cfsantos@ufscar.br
[2] State University of Sao Paulo - UNESP, Sao Paulo, Brazil
thierrypin@gmail.com, joao.papa@unesp.br
[3] Cenpes, Petróleo Brasileiro S.A. – Petrobras, Rio de Janeiro - RJ, Brazil
colombo.danilo@petrobras.com.br

Abstract. Most Convolutional Neural Networks make use of subsampling layers to reduce dimensionality and keep only the most essential information, besides turning the model more robust to rotation and translation variations. One of the most common sampling methods is the one who keeps only the maximum value in a given region, known as max-pooling. In this study, we provide pieces of evidence that, by removing this subsampling layer and changing the stride of the convolution layer, one can obtain comparable results but much faster. Results on the gait recognition task show the robustness of the proposed approach, as well as its statistical similarity to other pooling methods.

Keywords: Convolutional Neural Networks · Deep learning · Gait recognition

1 Introduction

Sub-sampling layers, known as pooling, perform two essential tasks on Convolutional Neural Networks (CNN): (i) to reduce the number of hyperparameters, thus decreasing the computational cost for training and inference; and (ii) to hold a certain degree of space invariance by keeping the most relevant information. Deep learning techniques have achieved state-of-the-art results on image processing tasks since 2010. Image classification and localization competitions, such as ImageNET Large Scale Visual Recognition Challenge (ILSVRC) [22] and COCO (Common Objects in Context) [15], comprise such neural models in their top results mostly. Inception-V4 [24] and ResNET [6], for instance, achieved outstanding results in image classification tasks. Their basic structure has been used in several other works by adopting transfer learning techniques [3,4,11,20].

João Paulo Papa–The authors acknowledge FAPESP grants 2013/07375-0, 2014/12236-1, 2016/06441-7, and 2017/25908-6, CNPq grants 429003/2018-8, 427968/2018-6 and 307066/2017-7, as well as Petrobras research grant 2017/00285-6.

I. Nyström et al. (Eds.): CIARP 2019, LNCS 11896, pp. 751–760, 2019.
https://doi.org/10.1007/978-3-030-33904-3_71

However, a considerable drawback of these networks concerns the computational cost for both training and inference, taking several days (or even weeks) to achieve the desired results. Therefore, any gain on speed is always welcomed in such models. This work aimed to introduce a more efficient way to reduce the number of parameters and still to keep the spatial invariance expected in CNN-based models. The idea is to replace pooling layers by 2D convolutions with stride as of two. Such modification keeps the average accuracy in different networks, with the boost in both training and inference time.

The remainder of this work is organized as follows: Sect. 2 describes several types of sub-sampling approaches, and Sect. 3 presents the proposed approach. Sections 4 and 5 discuss the methodology and the experiments, respectively. Finally, Sect. 6 states conclusions and future works[1].

2 Related Works

Convolutional Neural Networks were designed based on human visual cortex [13]. In short, such a brain region has two main types of cells: (i) simple cells, which are computationally emulated by the CNN kernels; and (ii) complex cells, that can be found either in the primary visual cortex [7], secondary visual cortex, and the Broadman area 19 of the human brain [9]. The former cells are allocated in the primary visual cortex, and such structures respond mainly to edges and bars [8]. The former cells respond both to edges and gradings, like a simple cell, but also to spatial invariance. It means that such cells react to light patterns in a large receptive field on a given orientation.

Based on this biological information, LeCun et al. [13] developed the first successful CNN model. Its structure consists of a total of seven layers: two pairs of convolutions followed by an average pooling, two multi-layer perceptrons layer, and a final layer responsible for classification. Roughly speaking, a CNN uses pooling since its beginning.

Max-pooling was first proposed in 2011 [17] as a solution for gesture recognition problems. Since then, several works claim that such operation is the best sub-sampling rule for a CNN. However, some other rules, such as Global Averaging Pooling [14], may also be applied in other circumstances: in this specific case, it was designed to replace a multi-layered perceptron network in the final layers of a CNN since it tries to impose correspondences between feature maps and categories. Another sub-sampling approach is a forced concatenation of information from *MaxPooling* combined with the convolution of stride two. The work of Romera et al. [21], for instance, aimed at performing real-time pixel-level segmentation using such paradigm, achieving near state-of-the-art segmentation results.

Sometimes, data sub-sampling is not desired because spatial information is quite important, and any loss could affect the results. DeepMind claims, on its reinforcement learning work [16], that any kind of pooling could remove relevant

[1] The source code is available at https://github.com/thierrypin/gei-pool.

spatial information in several games so that the CNN used in their work consists only on convolutional and perceptron layers. Therefore, such arguments suggest it may be necessary to develop new pooling techniques in order to improve results on several problems.

In this work, we proposed GEINet, a deep network for the problem of gait recognition that does not contain any pooling layer. Besides, we also showed that the lack of such a layer could provide satisfactory results, but pretty much faster.

3 Proposed Approach

The main goal of this work is to find out the best neural structure in order to perform gait recognition successfully. Proposed by Han and Bhanu [5], the Gait Energy Image (GEI) approach can be used to classify or identify a given individual. Such technique consists of an average of pictures from a person in a given activity, such as walking or jogging. Roughly speaking, it can be understood as a heatmap indicating what the most frequent positions assumed by a person are. Figure 1 depicts some examples of images generated by the GEI approach.

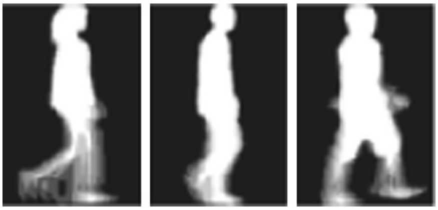

Fig. 1. Example of a GEI image for three different people. Image extracted from the "OU-ISIR Gait Database, Large Population Dataset (OULP)" [10].

State-of-the-art GEI classification results were achieved by Shiraga et al. [23], which proposed three other architectures to identify people from their gait images. The original network is straightforward, consisting of two blocks with a convolutional step (18 7×7 and 45 5×5 kernels), a 2×2 max-pooling, as well as local response normalization [12]. Following the convolutions, are two fully-connected layers of size $1,024$ and 956 (number of classes). All layer outputs are activated with ReLU, except for the last one, which is activated with the well-known *softmax* function.

In this paper, we proposed three other architectures for comparison purposes:

1. A re-trained GEINet structure composed of two sets of layers of convolution, pooling, and Local Response Normalization (LRN) [12]. Such layers are then followed by two multilayer perceptrons and finally by a *softmax* for baseline purposes;

2. A similar model, but removing the pooling layer, and changing the convolution stride from one to two (GEINet no-pool); and
3. A third model based on the first one, but replacing the pooling layer for a convolution layer of stride two, acting as a dimensionality reducer. This model doubles the number of convolution layers in comparison to the other two (Double-conv).

Figure 2 depicts the architectures of the neural networks proposed in this work.

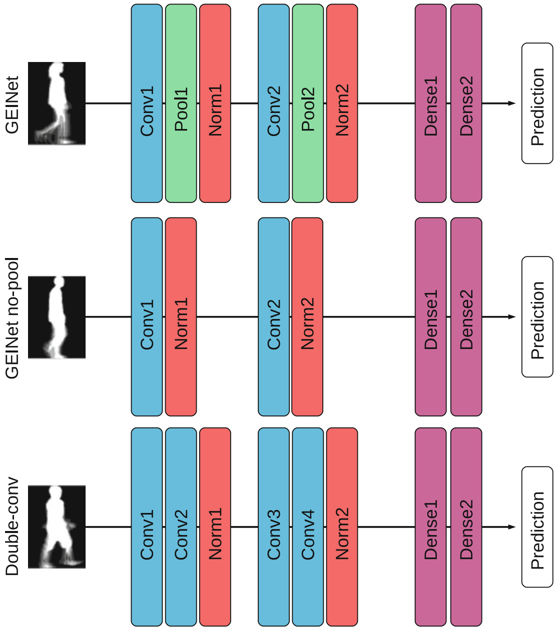

Fig. 2. Architecture of the neural networks proposed in this work.

We followed the protocol described by Shiraga et al. [23] to construct the energy images, which consists of taking four consecutive video silhouette masks to further obtaining their pixel-wise averages.

4 Methodology

In this section, we described the methodology employed to validate the robustness of the proposed approach. The equipment used in the paper was an Intel Xeon Bronze® 3104 CPU with 6 cores (12 threads), 1.70 GHz, 96 GB RAM 2666 Mhz, and GPU Nvidia Tesla P4 8 GB. The framework MXNet [1] was used for the neural network architecture implementation. We provided a better description of data sets used, models, and the evaluation protocol in the following subsections.

4.1 Data Set

We considered the "OU-ISIR Gait Database, Large Population Dataset (OULP)" [10], which consists of silhouettes from $3,961$ people from several ages, size, and gender, walking on a controlled environment. Data have been collected since March 2009 through outreach activity events in Japan and recorded at 30 frames per second, from four different angles: 55, 65, 75 and 85 degrees. The original images have a resolution of 640 × 480 pixels, but the silhouettes were further cropped originating another set of image with a resolution of 88 × 128 pixels. In this work, we resized the images to a resolution of 44 × 64 pixels for the sake of computational load.

4.2 Evaluation Protocol

We performed the cross-validation protocol described by Iwama et al. [10]. The dataset is divided into five subgroups of $1,912$ people each, and each subset i is further divided into two equal parts of 956 individuals, hereinafter called g_{i1} and g_{i2}, respectively, $\forall i = 1, 2, \ldots, 5$. The former group ($g_{i1}$) is used for feature extraction purposes using the proposed approaches and baseline, and the latter set (g_{i2}) is employed for the classification step. Each subset is further divided in half, i.e., $g_{i1} = g_{i1}^T \cup g_{i1}^V$ and $g_{i2} = g_{i2}^T \cup g_{i2}^V$, where g_{ij}^T and g_{ij}^V stand for training and validating sets, respectively, $\forall j = 1, 2$. In this work, we opted to use two fast and parameterless techniques for the classification step: the well-known nearest neighbor (NN) [2] and the Optimum-Path Forest (OPF) [18,19][2]. Figure 3 depicts the aforementioned protocol.

As mentioned earlier, the dataset provides four camera angles: 55°, 65°, 75°, and 85°. Therefore, we opted to use a cross-angle methodology, i.e., we used a given angle for training purposes and all angles to evaluate the models. Each video contains between 15 and 45 frames, but we used only 4 to build the gait energy images[3]. To train the neural networks, in each batch iteration, we selected four random contiguous frames. For evaluation purposes, we divided the videos into consecutive non-overlapping clips and further classified each. The final prediction is the mode of all predictions in the sequence.

Since the networks are trained with a single video from each subject, we employed data augmentation to improve training diversity. For this purpose, we employed four image transformations, each with 50% chance of occurring independently: horizontal flip, Gaussian noise with zero mean and standard deviation as of 0.02, as well as random vertical and horizontal black stripes of width 3. Additionally, the random temporal cropping step functions as augmentation. Lastly, due to the low number of videos and the high number of possible variations in the augmentation step, we trained the networks on $12,500$ epochs. In addition, we considered three measurements: (i) training and (ii) classification

[2] We used the Python OPF implementation available at https://github.com/marcoscleison/PyOPF.

[3] We observed that only four images were enough to obtain a reasonable energy image.

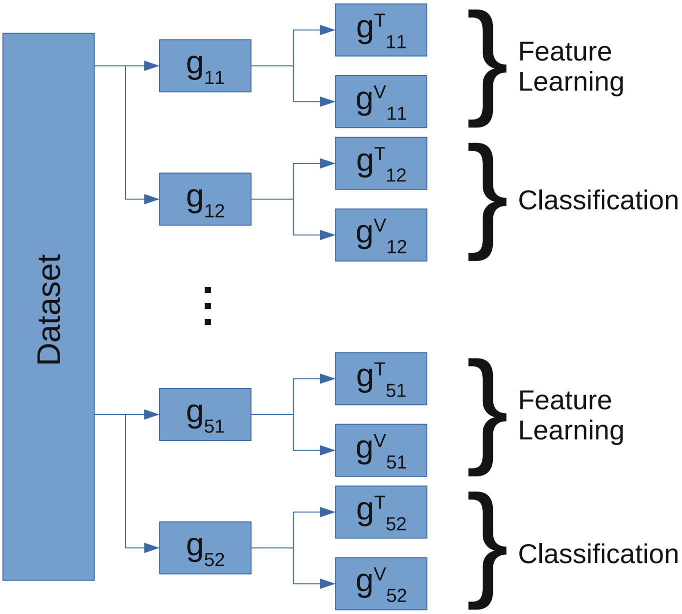

Fig. 3. Protocol adopted in the work, as described by Iwama et al. [10]. The dataset is divided into 5, and each part is further subdivided twice: one is used for feature learning and the other for classification. Then the parts are switched, so that there are a total of 10 evaluation steps.

times, and (iii) accuracy. Notice we used the Wilcoxon signed-rank test [25] for the statistical analysis of each measurement.

5 Results

In this section, we presented the experimental results and discussion. We showed that replacing the pooling layer by a larger convolutional stride is sufficient to obtain a good trade-off between computational load and accuracy. As aforementioned, in this paper we evaluated three models and compared their performance.

5.1 Accuracy

We evaluated how the models perform when predicting with different camera angles. All the training step was performed on a single camera angle and the same for the classifier. Therefore, the idea is to predict gaits from all four viewpoints. Tables 1 and 2 depict the accuracy results using NN and OPF, respectively. The results concern the average from all five folds, as described in Sect. 4.2. It is worth noticing that, the closer the test angle is to 90°, the better the overall accuracy is, i.e., when the camera records the actor from the side view. As expected, the accuracies tend to be higher when the train and test angles are the same.

Table 1. Mean accuracies using NN classifier.

Train angle	Method	Test angle			
		55	65	75	85
55	GEINet	88.77	88.56	87.51	88.35
	GEINet no-pool	87.82	87.26	86.05	87.34
	Double-conv	88.28	88.18	86.86	86.51
65	GEINet	85.61	89.52	90.27	90.88
	GEINet no-pool	83.56	88.41	88.87	89.77
	Double-conv	84.21	89.23	89.06	90.00
75	GEINet	79.71	86.80	90.29	91.59
	GEINet no-pool	79.54	87.13	90.15	91.78
	Double-conv	80.00	87.49	89.83	91.11
85	GEINet	72.57	78.85	87.07	91.42
	GEINet no-pool	73.37	79.48	87.15	91.17
	Double-conv	75.71	80.48	87.59	91.44

Table 2. Mean accuracies using OPF classifier.

Train angle	Method	Test angle			
		55	65	75	85
55	GEINet	88.14	88.01	87.02	88.14
	GEINet no-pool	87.36	86.99	85.67	87.07
	Double-conv	87.55	87.89	86.34	86.19
65	GEINet	85.12	89.08	89.81	90.48
	GEINet no-pool	82.80	88.08	88.26	89.41
	Double-conv	83.62	88.70	88.85	89.52
75	GEINet	79.27	86.09	89.79	91.28
	GEINet no-pool	79.16	86.67	89.67	91.42
	Double-conv	79.27	86.57	89.54	90.61
85	GEINet	71.99	78.10	86.36	90.86
	GEINet no-pool	72.87	78.81	86.97	90.86
	Double-conv	75.17	79.95	87.05	91.05

When replacing the pooling layer from GEINet by a stride in its convolution layer, the accuracy results go down marginally – around 1%. Besides, Wilcoxon test returned a p-value around to 10^{-7}, indicating they probably do not diverge. Trading the pooling step by a new convolutional layer with stride as of 2 results in slightly better results, but still not quite better than the original model. The Wilcoxon test outputted a p-value as of 0.102, indicating that their distribution might be similar as well.

5.2 Execution Time

Since the protocol employed in this paper establishes ten runs, and the models were trained once for each angle, each model has 40-time measurements. Therefore, all results presented in this section correspond to the average of all runs.

Table 3 presents the network training and inference times. Although the non-pooling model achieved slightly smaller accuracies than the original one, its training time is considerably lower. The reduction from $3,753$ seconds to $3,322$ corresponds to a gain of 11.5%, while such gain was 8.3% for inference purposes.

Table 3. Training and inference times: replacing the pooling layer by a convolutional stride resulted in considerably faster training time.

Training time			
Model	Per epoch (s)	Total (s)	Inference Time (s)
GEINet	0.300	3,753.71	0.108
GEINet no pool	0.266	3,322.18	0.099
Double-conv	0.320	4,004.93	0.109

6 Conclusion and Future Works

In this work, we introduced two variants of a simple but efficient model for gait recognition purposes (*GEINet*): one replaces the pooling layers by a convolutional stride (*GEINet no-pool*), and the other replaces the pooling layers by a convolutional layer with stride (*double-conv*). We showed the non-pooling version achieved slightly smaller accuracies than GEINet, but with a considerable speed-up (11.5%). On the other hand, the double-conv model ran 6.3% slower without any perceptible gain in accuracy. Regarding future works, we intend to use GEI to identify people directly from the video streams. Besides, different activation functions shall be investigated too.

References

1. Chen, T., et al.: Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. CoRR abs/1512.01274 (2015). http://arxiv.org/abs/1512.01274
2. Cover, T.M., Hart, P.E., et al.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**(1), 21–27 (1967)
3. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115 (2017)

4. Habibzadeh, M., Jannesari, M., Rezaei, Z., Baharvand, H., Totonchi, M.: Automatic white blood cell classification using pre-trained deep learning models: Resnet and inception. In: Tenth International Conference on Machine Vision (ICMV 2017), International Society for Optics and Photonics, vol. 10696, p. 1069612 (2018)
5. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Trans. Pattern Anal. Mach. Intell. **28**(2), 316–322 (2006)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015). http://arxiv.org/abs/1512.03385
7. Hubel, D., Wiesel, T.: Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. J. Physiol. **160**, 106–154 (1962)
8. Hubel, D.H., Wiesel, T.N.: Receptive fields of single neurons in the cat's striate cortex. J. Physiol. **148**, 574–591 (1959)
9. Hubel, D.H., Wiesel, T.N.: Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. J. Neurophysiol. **28**(2), 229–289 (1965)
10. Iwama, H., Okumura, M., Makihara, Y., Yagi, Y.: The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. IEEE Trans. Inf. Forensics Secur. **7**(5), 1511–1521 (2012)
11. Kong, B., Wang, X., Li, Z., Song, Q., Zhang, S.: Cancer metastasis detection via spatially structured deep network. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., Shen, D. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 236–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_19
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
13. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
14. Lin, M., Chen, Q., Yan, S.: Network in network. CoRR abs/1312.4400 (2013). http://arxiv.org/abs/1312.4400
15. Lin, T., et al.: Microsoft COCO: common objects in context. CoRR abs/1405.0312 (2014). http://arxiv.org/abs/1405.0312
16. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015). https://doi.org/10.1038/nature14236
17. Nagi, J., et al.: Max-pooling convolutional neural networks for vision-based hand gesture recognition. In: 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pp. 342–347. IEEE (2011)
18. Papa, J.P., Falcão, A.X., Suzuki, C.T.N.: Supervised pattern classification based on optimum-path forest. Int. J. Imaging Syst. Technol. **19**(2), 120–131 (2009). https://doi.org/10.1002/ima.v19:2
19. Papa, J.P., Falcão, A.X., Albuquerque, V.H.C., Tavares, J.M.R.S.: Efficient supervised optimum-path forest classification for large datasets. Pattern Recogn. **45**(1), 512–520 (2012)
20. Rakhlin, A., Shvets, A., Iglovikov, V., Kalinin, A.A.: Deep convolutional neural networks for breast cancer histology image analysis. In: Campilho, A., Karray, F., ter Haar Romeny, B. (eds.) ICIAR 2018. LNCS, vol. 10882, pp. 737–744. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93000-8_83
21. Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R.: Efficient convnet for real-time semantic segmentation. In: IEEE Intelligent Vehicles Symposium (IV), pp. 1789–1794 (2017)

22. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
23. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Geinet: view-invariant gait recognition using a convolutional neural network. In: 2016 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2016)
24. Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, inception-resnet and the impact of residual connections on learning. CoRR abs/1602.07261 (2016). http://arxiv.org/abs/1602.07261
25. Wilcoxon, F.: Individual comparisons by ranking methods. Biom. Bull. **1**(6), 80–83 (1945)