REALTIME PEDESTRIAN RECOGNITION USING SIAMESE NETWORK

Martin Rajnoha

Doctoral Degree Programme (2), FEEC BUT

E-mail: martin.rajnoha@vutbr.cz

Supervised by: Radim Burget

E-mail: burgetrm@feec.vutbr.cz

Abstract: Image similarity measuring has many various applications. Pedestrian recognition is one of them and for the security purposes it is basically required to run in real-time. This paper proposes a deep Siamese neural network architecture for pedestrian recognition that achieves 70.28% accuracy on the test set containing 20 persons. Prediction of the model is fast enough for real-time processing.

Keywords: surveillance, pedestrian, recognition, Siamese, deep learning

1 INTRODUCTION

Pedestrian recognition system can play significant role in our life. Such a system could help in many areas, but probably the most important one is security. Imagine how many people are walking through the airports, bus stations shopping mall every moment. Even all these public places have security cameras around they are basically controlled by human operator. Pedestrian recognition system could significantly help with almost impossible tasks for human. For example, in the case of robbery in some store in shopping mall. The store has an image of suspect from their security camera. They can send an image to security operator or make a call and provide description of suspect. It is almost impossible for human to find suspect among the people in the shopping mall. There can be plenty of cameras and hundreds of walking people. Pedestrian recognition system could solve this problem in seconds, maybe even milliseconds.

This paper deals with pedestrian recognition using deep Siamese convolutional neural network. One of the crucial point for recognition is also the processing time. Architecture of the Siamese network for pedestrian recognition was proposed and train using created dataset containing 46 persons and 1230 photos. The rest of the paper is structured as follows: Next section – methodology theoretically describes basics of convolutional neural networks and Siamese networks. Dataset for training Siamese network is described in this part as well and the experiment part presents the proposed method for pedestrian recognition and results. The last section concludes the paper.

2 METHODOLOGY

To make pedestrian recognition works as a whole system, first there is need to have some detection system. In this work is used a detection method described in [1]. From individual detections containing only one person on the image it is possible to do recognition. There are many techniques for image processing but in the recent years Convolution Neural Networks (CNN) are the most widely used for this area. CNNs have become a standard approach for problems dealing with image classification [1],[2].

Image classification is one of the simplest tasks regarding image processing using deep learning. General architecture for classification consists of few Convolutional layers with ReLU activation function and each followed by Max-pooling layer. Next there is a Flatten layer and after this layer

are few Fully Connected layers, also called Dense layers. Usually between them are Dropout layers for reducing of overfitting. Last layer is Dense layer, sometimes called Activation or Classification layer and its contains as much neurons as number of classes [3]. Example of the basic convolutional neural network used for classification is in fig. 1. This network was used for Handwriting recognition [4].

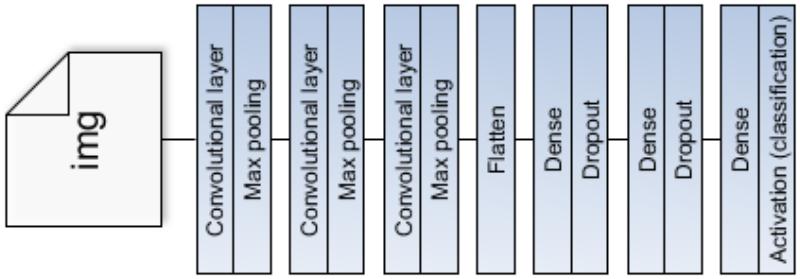

Figure 1: Example of convolutional neural network with basic architecture for classification [4].

Principle of the feature extraction from image data is very similar for classification and recognition. Recognition process compares two images and estimates their similarity. This is not possible to achieve with basic architecture of CNN, because in recognition case there are two images on the input to the network. Siamese neural networks are used for these purposes.

2.1 SIAMESE NETWORK

Siamese neural nets were firstly introduced to solve image matching problem in the early 1990s [5]. Siamese neural network consists of two identical neural networks, each has one image as an input. Because of networks are the same, they compute the same function and thus they make exactly same feature extraction for both images. Basic idea of Siamese neural network architecture is presented in fig. 2.

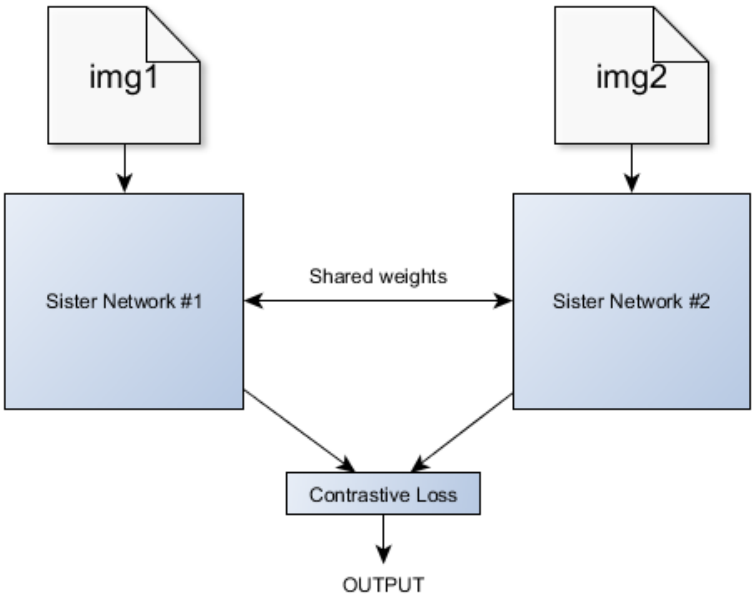

Figure 2: Example of basic idea of Siamese neural network.

Instead of learning to classify its inputs, it learns the similarity between inputs. Actually, it measures difference between images using contrastive loss function. This function is Distance-based and can use for example Euclidean distance. Definition of contrastive loss function [6]: „*A*

contrastive loss function is employed to learn the parameters W of a parameterized function GW, in such a way that neighbors are pulled together and non-neighbors are pushed apart. Prior knowledge can be used to identify the neighbors for each training data point."

2.2 DATA DESCRIPTION

For training and testing of system for pedestrian recognition there is necessary to have a dataset containing few persons with multiple photos. Dataset in our work were created from video used for 'i-Lids bag and vehicle detection challenge' on IEEE International Conference on Advanced Video and Signal based Surveillance, 2007 [7]. Using detection method described in [1] pedestrians were detected every 2 seconds and detections were stored. After that all the detections had to be sorted in the way that one pedestrian contains only photos itself. Dataset contains 56 pedestrians with overall 1179 samples distributed in pedestrians. The dataset was divided into two sets using for training of a model and for testing of performance of the model.

Because of Siamese networks has on input two images, the form of data for training has to be in shape: [image1, image2, label], where label defines whether these two images are same person or not. A datagenerator was created for this purpose. Datagenerator generates a pair of images and creates label according to sources of the images. For training process is important to have both positive and negative samples. Process of generating balanced dataset is as follows:

- **Positive samples** are created as all possible combination of images for all persons. There are 36 persons with overall 1029 photos in training set. Datagenerator automatically created 28230 positive pairs without repeating of images.
- **Negative samples** are created from two random selected persons and their random selected images. Pairs of images do not repeat. Datagenerator generates negative samples until it reaches number of positive samples. This is important to have a balanced dataset.
- **Combination** of positive and negative samples, interleaved one by one, the training dataset consists of 56460 samples – pairs of images, half positive (same person) and half negative (different persons).

The testing set is created in the same way and contains 562 pairs from 20 persons with overall 150 images. Testing set is used only for evaluation purposes, it does not affect training process. Training set is split in two set called training (30000 samples) and validation (26460 samples). Positive and negative pair is shown in fig. 3.

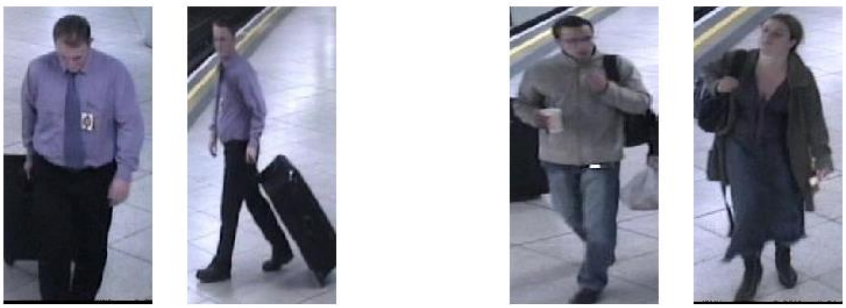

Figure 3: Example of pairs of images generated by datagenerator for training and testing of a model (a) positive pair (same person), (b) negative pair (different persons).

2.3 EXPERIMENT

Proposed deep Siamese convolutional neural network is represented in fig. 4. As an input to proposed architecture of Siamese neural network are two images with 64 x 128 pixels size. At the end of the network is similarity measurement layer with contrastive loss function. In this work we used Euclidean distance for measuring similarity of both feature vectors extracted by identical Siamese networks. If there are same images on the input, both parts of network will generate same vectors.

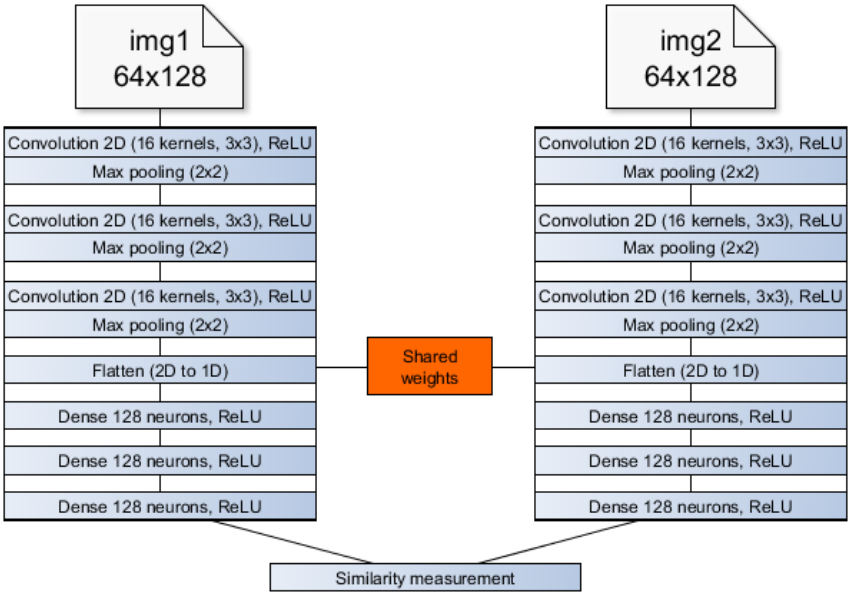

Figure 4: Proposed Siamese neural network architecture with shared weights.

Within the experiment many architectures and different parameters settings were tried. For example number of Convolutional layers, number of Dense layers, kernel sizes, number of kernels, number of neurons etc. Keras [8] and Tensorflow [9] frameworks were used to compile and train the network. These frameworks allow acceleration on GPU.

Performance of the model is 81.66% accuracy on the training set and 70.28% accuracy on the testing set which does not affect training process. Prediction of the model returns a value representing a difference between two input images. Dataset for training and testing is created in classification way, it means two images with same person got label 1 (positive) without distinguishing their similarity. Same label got two images with almost identical position, size, etc. of the person and the same label got two images with different posing, size, lighting of the same person. For evaluation of the accuracy had to be set some threshold value which determines if predicted value represents same person or not. In this case we used threshold value 0.5 and values lower than 0.5 are considered as same person and greater than 0.5 are considered as different persons.

One of the important part of the pedestrian recognition is also computational time. Using this neural network is possible to achieve 16,92ms of average prediction time. It includes pre-processing such as resizing and normalization of both images. The measuring of average time prediction was evaluated at the computer with processor Intel® Core™ i7-6700, 3.40GHz, 64-bit Windows 10, 32 GB RAM and the model was loaded and runs using graphic card Nvidia GeForce 1080 Ti.

3 CONCLUSION

This paper aimed about challenging task in the image processing area - real-time pedestrian recognition. For this purpose, the dataset with 56 persons and 1179 images was created. The architecture of deep Siamese neural network was proposed to solve this task. Average prediction time of the

network for two images is 16.92ms. It was proved that Siamese neural networks are suitable for similarity measuring tasks, especially for image signals.

The performance of the proposed architecture is 70.28% of accuracy on the independent testing dataset. This accuracy is relatively low but there are few possible explanations for it. The first is size of the dataset. The model has tendency to overfit because of little amount of training data. On the other hand, by simplifying the proposed architecture of Siamese neural network would not be possible to generalize and learn network for this task. Increasing of training dataset should cause increasing of performance model and avoid overfitting.

The second explanation deals with computing of accuracy. Testing set is created in classification way, it means there are labels 0 (negative) for different persons and 1 (positive) for same person. Instead of that, the network is trained to estimate difference between images. It means there must be some threshold value to define whether images are from same person or not.

ACKNOWLEDGEMENT

Research described in this paper was financed by NSP LO1401 and by MVCR VI20172019086. For the research, infrastructure of the SIX Centre was used.

REFERENCES

[1] RAJNOHA, M.; POVODA, L.; MAŠEK, J.; BURGET, R.; DUTTA, M. Pedestrian Detection from Low Resolution Public Cameras in the Wild. In 5th International Conference on Signal Processing and Integrated Networks (SPIN), New Delhi, India. 2018.

[2] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[3] F. Chollet, "Building powerfull image classification models using very little data," The Keras Blog [online], June, 2016, [cited 15.03. 2018]. Available from: https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html

[4] M. Rajnoha, R. Burget, and M. K. Dutta, "Handwriting comenia script recognition with convolutional neural network," 2017 40th International Conference on Telecommunications and Signal Processing (TSP), pp.775–779, Jul. 2017.

[5] G. Koch, R. Zemel, R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In ICML Deep Learning Workshop 2015 (Vol. 2).

[6] R. Hadsell, S. Chopr, Y. LeCun. Dimensionality reduction by learning an invariant mapping. In Computer vision and pattern recognition, 2006 IEEE computer society conference on 2006 (Vol. 2, pp. 1735-1742). IEEE.

[7] AVSS 2007, IEEE Internation Conference on Advanced Video and Signal based Surveillance, London (United Kingdom), September 2007, [online], [cited 15.3. 2018]. Available from: http://www.eecs.qmul.ac.uk/ ~andrea/avss2007_d.html

[8] F. Chollet, Keras: Deep Learning library for Theano and TensorFlow, 2015. Available: https://keras.io/

[9] TensorFlow ™, An open-source machine learning framework for everyone. Available: https://tensorflow.org