REVIEW OF BACHELOR THESIS

AUTHOR: ALENA MORAVOVÁ
TITLE: TEXT RECOGNITION IN IMAGES USING RECCURENT NEURAL NETWORKS
REVIEWER: BORIS FLACH

The bachelor thesis presented by Alena Moravová considers the task of text recognition in images from natural scenes. It is assumed that the rectangular input images represent single words, cropped from natural scenes by some text detection algorithm. The author proposes a combination of two neural networks together with an elementwise independent probabilistic sequence model for inferring the character string represented in the image.

The first part of the model is a deep convolutional network that maps rectangular receptive fields of the input image into appropriate feature vectors. When applied for sliding windows, this results in a sequence of feature vectors. The latter are then interpreted by a recurrent neural network, resulting in a sequence of character label probability vectors of fixed length. The last inference step infers a character string from these probabilities.

In order to learn all model parameters jointly from a training set of annotated examples, it is necessary to perform the following steps.
(1) To compute the label probabilities for each window by a "forward-backward" algorithm similar to computing marginal probabilities for HMMs.
(2) To apply backpropagation through time to learn the weights of the recurrent network.
(3) Learn the weights of the convolutional network by backpropagation.

Starting from an introduction to neural networks, the author describes all necessary concepts in the first two technical chapters of her thesis. The final chapter presents experiments on artificial data (used for training) and two datasets obtained from real images (used for testing). Special focus was put on questions as correct scaling, case sensitive recognition and others.

The thesis is well structured and almost everywhere clearly written. However, the balance could be improved. Too much space is spent in my view on well known basics of neural networks. On the other hand, the advanced concepts presented in the thesis would have deserved a more detailed description including a more elaborate notation in some places.

A list of further comments and questions follows.

- The author refers to an universal approximation property, which claims that every function can be approximated by a network with one hidden layer (p. 6). This should be made more precise. On the other hand, on p. 10 the author says that deep networks can learn "better representations" of complex inputs. Is this a contradiction?

- Convolutional networks with ReLU activation functions and max-pooling layers represent non-differentiable mappings. Please explain, how to apply backpropagation for such mappings in a mathematically correct way.
- Please explain the claim (p. 29) that models with more classes, as e.g. the case sensitive recognition, will automatically lead to less good results?
- It would be natural to do the last inference step by choosing the most probable character string. Please explain, why instead you use the most probable label sequence (p. 22)?

Notwithstanding these remarks and questions, I consider the thesis as fully acceptable. My grading is “very good” (B).

Prague, 13.06.2016 Dr.rer.nat.habil. Boris Flach