CS-E4890 DEEP LEARNING exam 12.2.2018 **Kivinen, Laaksonen, Keurulainen**

Allowed equipment: pens, pencils, erasers, a non-programmable calculator

Question 1 [max. 9 points]: Explain the following concepts briefly but informatively (each with 30–50 words, with a mathematical definition and/or with an illustration):

(i) underfitting versus overfitting

(ii) deep belief networks

(iii) Hessian matrix

(iv) input data normalization

(v) linear factor models

(vi) dataset augmentation

(vii) weight decay

(viii) denoising autoencoders

(ix) saddle point

Question 2 [max. 6 points]: (i) Describe the stochastic gradient descent algorithm verbally and with an equation and an illustration. (ii) Similarly, describe how momentum is used in optimization. (iii) Also, describe the Adam optimization algorithm similarly (no illustration needed). (iv) Compare the pros and cons of the above three optimization methods.

Question 3 [max. 4 points]: (i) Explain what is meant with the parameters and hyperparameters of a deep neural network. (ii) Name some deep neural network hyperparameters and explain how changing their values can be expected to affect the effective capacity of the network model. (iii) Describe different approaches for selecting the optimal values for the hyperparameters of a deep neural network.

Question 4 [max. 4 points]: (i) Let's assume a four-layer feedforward network that is specified by the following equations: $z = f_3(y)$, $y = f_2(x)$, $x = f_1(w)$. Solve what is $z(w)$. (ii) Draw a forward computational graph of $z(w)$. (iii) Draw a computational graph that shows the symbol-to-symbol derivatives needed to compute $\frac{dz}{dw}$. Based on the derivatives in the graph, solve what is $\frac{dz}{dw}$. (iv) By substituting the following definitions: $f_3(t) = (1 + \exp(-t))^{-1}$, $f_2(t) = 3t - 4$, $f_1(t) = t^2$, compute the values of $z(w = 1)$ and $\frac{dz}{dw}(w = 1)$.

Question 5 [max. 4 points]: (i) Describe verbally and mathematically a representative variational autoencoder and a representative restricted Boltzmann machine. (ii) What kinds of general properties do they have in common and what are the main differences between them. (iii) What are their benefits and drawbacks when compared with each other, and in what kinds of applications one can use such models?

Question 6 [max. 6 points]: (i) Describe the general principle of parameter sharing and why it is useful in training deep neural networks. (ii) Give an example and explain how parameter sharing is used in a convolutional neural network (CNN). (iii) Similarly, give an example and explain how parameter sharing is used in a recurrent neural network (RNN). (iv) What different kinds of problems there may exist in training a CNN versus training an RNN?