UNIVERSITEIT TWENTE.

Applied Analysis Group
Department of Applied Mathematics
Dr. Christoph Brune

Exam: Deep Learning - From Theory to Practice (201800177)

Date: Tuesday, Feb 22, 2018
Time: 8:45-11:45

- The exam has 40 points in total (grading scheme below).
- An explanation to every answer is required. Answers can be short and focussed.
- You can make use of a calculator. No other additional material is allowed.

Good luck and success!

Exercise 1. (Backpropagation, MLP) [6pts]

(a) **[2pts]** *Describe shortly with words what backpropagation means? How is it related to the loss function of a neural network?*

(b) **[2pts]** Suppose you design a multilayer perceptron (MLP) for classification with the following architecture. It has a single hidden layer with the hard threshold activation function, i.e.

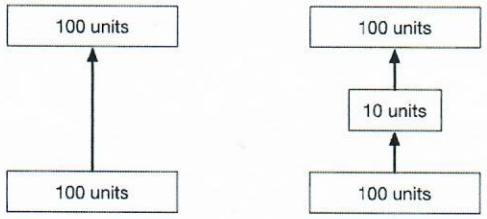
$$\sigma_{hthres}(x) := \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}.$$

The output layer uses the softmax activation function, i.e.

$$p_i = \sigma_{smax}(a_i) := \frac{e^{a_i}}{\sum_{k=1}^{N} e^{a_k}}$$

with cross-entropy loss $H(y,p) := -\sum_i y_i log(p_i)$.
What will go wrong if you try to train this network using gradient descent? Justify your answer in terms of the backpropagation rules.

(c) **[2pts]** Consider the following two MLPs, where all of the layers use linear activation functions.



Give two advantages of Network A over Network B, and two advantages of Network B over Network A.

Exercise 2. (Convolutional Neural Networks) [7pts]

(a) **[3pts]** Suppose you have a convolutional neural network (CNN) with the following architecture:

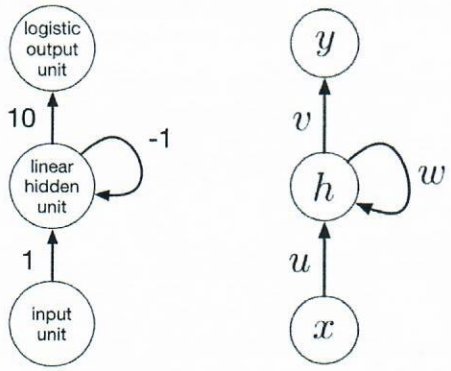
- The input is an RGB image of size 256×256.
- The first layer is a convolution layer with 32 feature maps and filters of size 3×3. It uses a stride of 1, so it has the same width and height as the original image. If needed, you may make assumptions on the padding.
- The next layer is a pooling layer with a stride of 2 (so it reduces the size of each dimension by a factor of 2) and pooling groups of size 3×3.

Determine the size of the receptive field for a single unit in the pooling layer. (i.e., determine the size of the region of the input image which influences the activation of that unit.) *You may assume the receptive field lies entirely within the image. Hint: you may want to draw a one-dimensional convolutional network to reason about this problem.*

(b) **[4pts]** *What is meant by the curse of dimensionality in machine learning? What are the three key properties of CNNs? Please explain them shortly.*

Exercise 3. (Recurrent Neural Networks) [10pts] Logistic function: $\sigma(z) := \frac{1}{1+e^{-z}}$

(a) **[3pts]** *Determine what the following recurrent neural network (RNN) computes.* More precisely, determine the function computed by the logistic output unit at the final time step; the other outputs are not important. All of the biases are 0. You may assume the inputs are integer valued and the length of the input sequence is even. The numbers u, v and w are weights to be multiplied with.



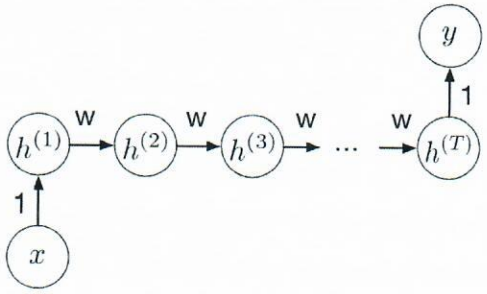
(b) **[3pts]** Consider a residual network built out of residual units, where the residual function is given by an MLP with one hidden layer:

$$z = W^{(1)}x + b^{(1)}$$
$$h = \phi(z)$$
$$y = x + W^{(2)}h$$

Provide a way of setting the weights and biases such that the derivatives (in computing the loss gradient) will not explode or vanish. Briefly explain your answer, but you do not need to provide a detailed derivation. What is the role of the activation function ϕ in this example regarding exploding or vanishing gradients?

(c) **[4pts]** Consider the following RNN, which has a scalar input at the first time step, makes a scalar prediction at the last time step, and uses a shifted logistic activation function:

$$\phi(z) := \sigma(z) - 0.5 \, .$$



Write the formula for the derivative of the loss $\frac{\partial L}{\partial h^{(t)}}$ *as a function of* $\frac{\partial L}{\partial h^{(t+1)}}$ *for* $t < T$. Hint: Use $z^{(t)}$ to denote the input to the activation function at time t. You may write your answer in terms of σ', i.e. you don't need to explicitly write out the derivative of σ.
Now suppose the input to the network is $x = 0$. Notice that then $h^{(t)} = 0$ for all t. *Based on your previous answer, determine the value* α *such that if* $w < \alpha$ *the gradient vanishes, while if* $w > \alpha$, *the gradient explodes.* You may use the fact, that $\sigma'(0) = \frac{1}{4}$.

Exercise 4. (AutoEncoders, Generative Networks) [9pts]

(a) **[3pts]** *Why is a basic AutoEncoder not good at generation of new data? How does a Variational AutoEncoder (VAE) overcome this problem?*

(b) **[3pts]** *Describe the main modeling idea behind a Generative Adversarial Network (GAN) in 5 sentences. How does the mathematical model look like?*

(c) **[3pts]** One of the most significant and widely discussed problems of GANs is *Mode Collapse. What does it mean? Do you know another challenge of GANs?*

Exercise 5. (Regularisation) [8pts]

(a) **[2pts]** *What does generalisation of a deep learning model mean? Describe it shortly with words.*

(b) **[6pts]** To improve the generalisation property of deep learning networks we introduced at least six different regularisation techniques. *Can you mention three and explain them?*

Grading scheme:

Ex 1.	(a) 2pt (b) 2pt (c) 2pt	**Ex 2.**	(a) 3pt (b) 4pt	**Ex 3.**	(a) 3pt (b) 3pt (c) 4pt	**Ex 4.**	(a) 3pt (b) 3pt (c) 3pt	**Ex 5.**	(a) 2pt (b) 6pt

Total: 40 points