UNIVERSITEIT
TWENTE.

Mathematics of Imaging & AI
Department of Applied Mathematics
Prof. Dr. Christoph Brune
Dr. Len Spek, Julian Suk, Dr. Nicolò Botteghi

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

Date: Thursday, Jan 25, 2024
Time: 13:45-15:45

- The exam has 24 points in total (grading scheme below).
- An explanation to every answer is required. Answers can be short and focused.
- No additional material or tools are allowed.

Good luck and success!

Exercise 1. (General concepts in Deep Learning) [6pt]

(a) **[1pt]** Which of the following activation functions can lead to vanishing gradients for large inputs in a shallow neural network?

(i) ReLU
(ii) Tanh
(iii) Leaky ReLU
(iv) Sin

(b) **[1pt]** What is the main goal of a convolutional neural network (CNN)?

(i) To model sequential data.
(ii) To cluster high-dimensional data.
(iii) To learn features from local regions of the input.
(iv) To create art by generating images.

(c) **[1pt]** Which regularization method can be interpreted as placing a probabilistic prior on the neural network's weights, encouraging the network to learn a sparse representation?

(i) L^2-regularisation.
(ii) Early Stopping
(iii) Using an ensemble of models
(iv) Dropout

(d) **[1pt]** For a general graph convolutional layer, which property always holds?

(i) Translation equivariance
(ii) Translation invariance
(iii) Permutation equivariance
(iv) Permutation invariance

(e) **[1pt]** In deep Q learning, what is modelled using a neural network?

(i) Value function

(ii) State

(iii) Past rewards

(iv) Transition probabilities

(f) **[1pt]** When do we say that an AI model is well aligned?

(i) When the training loss is zero.

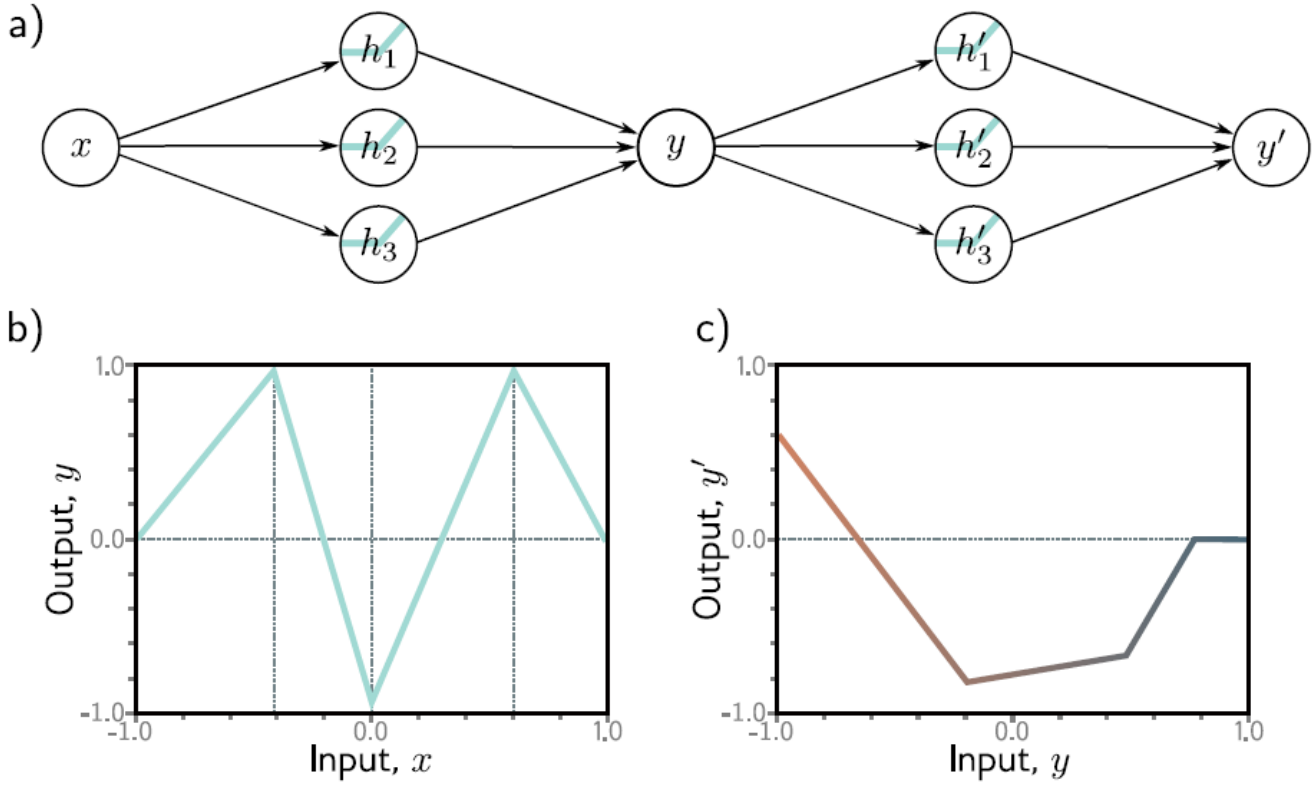
(ii) When the test loss is zero.

(iii) When the model is robust against adversarial attacks.

(iv) When the behaviour of the model corresponds to the intended goals of the designer.

Exercise 2. (Deep Neural Networks) [4pt]

(a) **[1pt]** Consider composing the two neural networks to get the network in figure (a). This first layer is plotted in figure (b), and the second layer is plotted in figure (c). Draw a plot of the relationship between the input x and output y' for $x \in [-1, 1]$.



(b) **[1pt]** Name 4 hyperparameters that are needed to train a multi-layer perceptron with a squared error loss function and weight decay, using batched stochastic gradient descent.

(c) **[2pt]** Consider a network with $D_i = 1$ input neurons, $D_o = 1$ output neuron, and $K = 10$ layers, with $D = 10$ hidden units in each. The network contains no biases. How would the number of weights increase more if we increased the depth by one? And how if we increased the width by one?

Exercise 3. (Residual Neural Networks) [7pt] We have time series data $x(t^1), \cdots, x(t^N) \in \mathbb{R}^d$ and want to learn a scalar feature $y \in \mathbb{R}$. There are many different architectures one could use.

One architecture is a residual neural network, with a hidden neuron $h \in \mathbb{R}^D$, given by

$$\begin{aligned} h^i &= h^{i-1} + \sigma(Wx(t^i) + Vh^{i-1}) \\ y &= Uh^N \end{aligned} \tag{1}$$

Here σ is the ReLU function.

(a) [**1pt**] Give the dimensions of the weight matrices W, V, U.

(b) [**1pt**] What is the gradient of y with respect to the jth element of the last hidden neuron h^N, i.e. $\frac{dy}{dh_j^N}$,

(c) [**2pt**] Write a recursive formula for the gradient of y with respect to the jth element of the $(i-1)$th hidden neuron h^i, $\frac{dy}{dh_j^{i-1}}$. The right-hand side should contain the terms like $\frac{dy}{dh_k^i}$.

(d) [**1pt**] For this architecture, which is more likely to occur: Vanishing gradients or exploding gradients? Use the gradient of question (c) in your answer.

(e) [**2pt**] A transformer is another architecture designed for time series. The main component is the attention mechanism $a[x(t^m), x(t^n)]$. Given keys k_m, queries q_n, weights Ω_k, Ω_q, biases β_k, β_q and the softmax activation function, how is the attention $a[x(t^m), x(t^n)]$ defined?

Exercise 4. (Generative Models) [7pt]

(a) [**2pt**] Sketch the architecture of a Variational Autoencoder (VAE) and mention a key added value compared to a basic Autoencoder.

(b) [**2pt**] Define in words what Normalizing Flows (NF) in the context of generative models are and give two key aspects where they differ from generative models like GANs or VAEs?

(c) [**2pt**] Consider transforming a uniform base density defined on $z \in [0, 1]$ using the function $x = \mathrm{f}[z] = z^3$. Find an expression for the transformed distribution $\Pr(x)$.

(d) [**1pt**] Explain shortly the challenge associated with training Generative Adversarial Networks (GAN), particularly focusing on the issue of mode collapse and training stability.

Grading scheme:

Ex 1.	(a) 1pt	**Ex 2.**	(a) 1pt	**Ex 3.**	(a) 1pt	**Ex 4.**	(a) 2pt
	(b) 1pt		(b) 1pt		(b) 1pt		(b) 2pt
	(c) 1pt		(c) 2pt		(c) 2pt		(c) 2pt
	(d) 1pt				(d) 1pt		(d) 1pt
	(e) 1pt				(e) 2pt		
	(f) 1pt						

Total: 24 points

Disclaimer: Some questions in this exam have been partially generated by GPT-4, based on previous exam questions and human input. All questions have been edited, critically reviewed and revised by the lecturers, to make sure that they correspond to the learning goals of the course.