Deep Learning Exam
WS 2023/24

Dr. Ozan Özdenizci

25th April 2024

First of all, I would like you to know that these notes were written 10 minutes after the exam finished, so they are as fresh as possible.

1 Architecture

1.1 a

There are two dense NN models with layer sizes $[D, 100, 100, K]$ and $[D, 100, 10, 10, 100, K]$. Both of them are using ReLu as activation. Which network will have more trainable parameters?

1.2 b

Explain why the sigmoidal activation function saturates in the hidden layers of the network. Why tanh will be better in this sense?

2 Optimization

2.1 a

What is the weight decay and how is it applicable to the MAP method?

2.2 b

There is a logistic function $y(x) = w_1 ln(x + w_2)$. Please, write an update rule for the w_2 parameter by using MSE loss function.

3 Transformers

3.1 a

Which role does the positional encoding play in attention transformer?

3.2 b

What is the computation of the attention transformer operation with D key/value dimension and L sequence length? Which troubles can it lead to?

3.3 c

Compare using Attention Transformer vs RNNs in for example machine translation.

4 Deep Generative Models

4.1 a

There is the loss function for the variational autoencoder.

$$L(\theta, \phi, X) = -D_{KL}(q_\phi(z|x)||p(z)) + E_{q_\phi(z|x)}[logp_\theta(x|z)]$$

Please explain what does each term means.

4.2 b

Here are the loss functions for the GAN model. Explain how do we update weights and where by using these losses.

$$E_{x\sim p_{data}(x)}[\log D(x)]E_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$
$$E_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$

Also explain mode collapse in GANs.

5 Practical task

We have M data samples (x_i, t_i) which are created by some function $y = (x, w)$ with noise probability $p(t|\mu, b) = \frac{1}{2b} exp^{-\frac{|\mu - t|}{b}}$. Please show that finding MLE for the w means minimizing the loss function $L = \sum_i^m |t_i - y(x, w)|$.