Generell info om digital campus eksamen - INF265 vår 22

This exam has 10 questions (You can see 11 questions in Inspera but Question 11 is a placeholder for points from projects). Each question is worth 5-10 points and the total number of points from the exam is 55 points.

Sheets for handwriting/drawing
On this exam it will be possible to attach hand-drawn sketches/illustrations or handwritten text to your digital exam answer. This is recommended for Question 10.

An exam question code will be available under each of the questions in the exam set. Ask the invigilator for drawing paper. The exam question code is unique for each question per student, so be sure to mark the sheet you have written or drawn on with the exam question number and the question code for the question you have answered on the sheet during examiniation.

In the 15 minutes after exam end time, you must shade the exam question code in the boxes under each digit in the question code and also fill out other requested information at the top of the page. Your candidate number can be found in the exam system.

Please ask an invigilator if you have trouble finding questions codes or your candidate number. When you have finished your exam, the sheets are to be submitted together, in the order they will be added to your answer paper, to the head invigilator in the venue.

1 Activation functions

Your friends Amanda, Bertha and Cecilie are considering using exotic activation functions in their neural network. They come up with the following suggestions:

Amanda: $f_A(z) = \max(0.3z, 3z)$
Bertha: $f_B(z) = 1.5z$
Cecilie: $f_C(z) = \cos z$

What are pros and cons of each suggestion?

Fill in your answer here

Maximum marks: 5

2 Initialisation strategies

Your friends Dag, Egil and Frode are training a feedforward neural network with one hidden layer. They have the following discussion.

Dag: I initialised the weights in my network by setting all of them zeros.
Egil: No, that is not a good strategy. I initialised the weights by setting all of them to ones.
Frode: No, you both are wrong. The best way to initialise weights is to set them all to 1/m where m is the number of neurons in the hidden layer.

Comment the strategies that your friends have chosen. What do you recommend them to do?

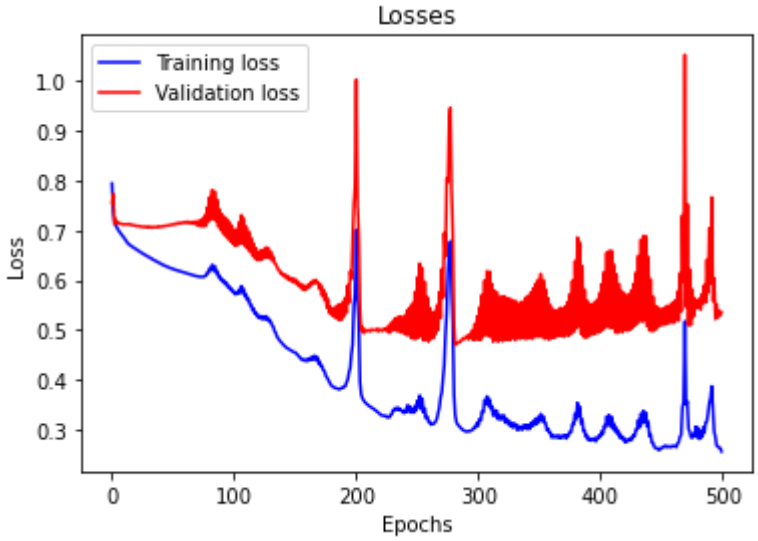
Fill in your answer here

Maximum marks: 5

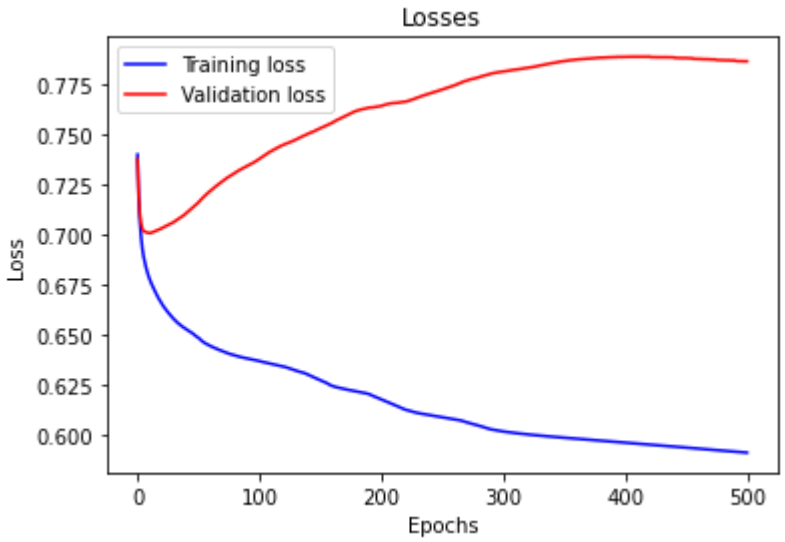
3 Training

Janne and Kari are both training neural networks for a classification task. They have plotted cross-entropy loss for both training and validation data after each epoch. You can see their plots below.

Janne:



Kari:



What can you infer from the plots? How would you advice Kari and Janne to improve their models?

Fill in your answer here

Maximum marks: 5

4 Forward propagation

Consider the following neural network:

$$Z^{[1]} = W^{[1]}x + b^{[1]}$$
$$A^{[1]} = relu(Z^{[1]})$$
$$\hat{y} = W^{[2]}A^{[1]} + b^{[2]}$$

with a two-dimensional input vector $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$

where

$W^{[1]} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, b^{[1]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, W^{[2]} = \begin{bmatrix} 1 & -1 \end{bmatrix}, b^{[2]} = 0.$

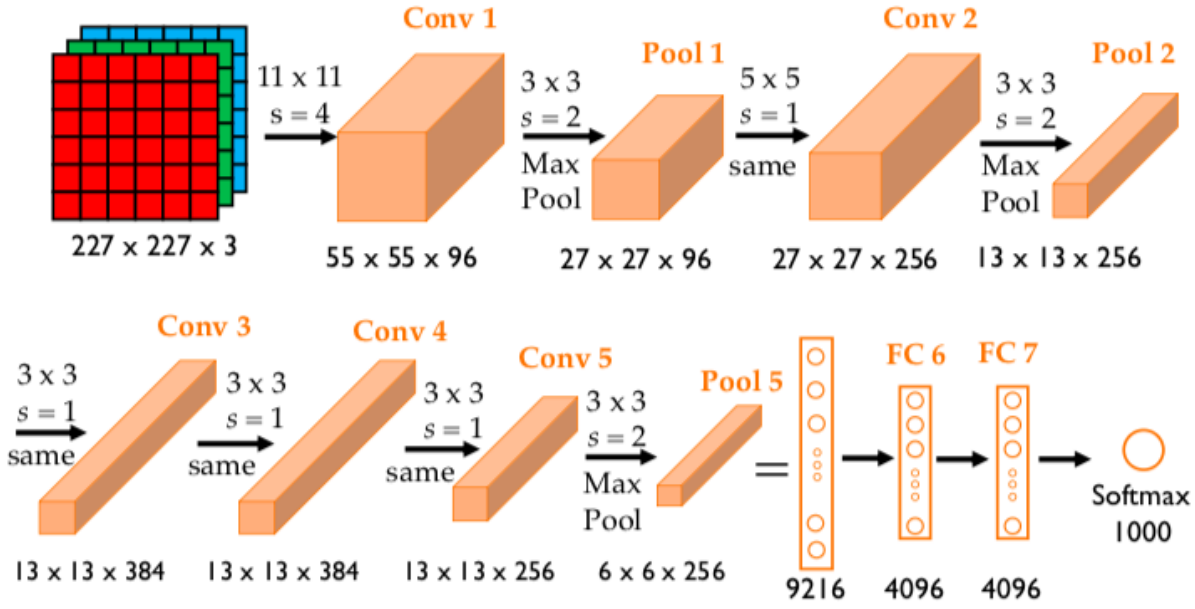
Which function does the network compute? Justify your answer.

Fill in your answer here

Maximum marks: 5

5 Alexnet

Suppose you are training a convolutional neural network based on the Alexnet architecture; see below.



(Image credit: Anh H. Reynolds)

This architecture has more than 62 million parameters. Suppose you want to modify the architecture to reduce the number of parameters. What kind of strategies can you take? Compare your strategies in terms of effectiveness. That is, which ones lead to large decrease in the number of parameters and which ones have a more moderate effect?

Fill in your answer here

Maximum marks: 5

6 Pooling layers

What are benefits of using pooling layers in convolutional neural networks?

Fill in your answer here

Maximum marks: 5

7 RNN

Give a short answer to the following questions:

1. What is a recurrent neural network and how does it differ from a feedforward neural network?
2. What is the vanishing gradient problem and why is it often mentioned in the context of recurrent neural networks?

Fill in your answer here

Maximum marks: 5

8 Recommender engine

Isak is designing a recommendation engine for an online store. The store wants to recommend products to its customers. The recommended products should be similar to the products that the customer has already bought.

The store has data about products that are purchased together. That is, if some customer has bought products x_1, $x_2$and x_3 at the same time, the data contain an item (x_1, x_2, x_3). Unfortunately, Isak has no access to any other data about the products. In other words, he does not have any features to decide similarity of products. Furthermore, there are too many different products to determine their similarity manually using the names of the products.

Isak has discussed with a consultant about the problem. The consultant proposed Isak to use machine learning. Unfortunately, Isak has forgotten details and can only remember that the consultant mentioned words *Word2vec* and *context*. Now, he asks you whether you can clarify what the consultant was suggesting.

Help Isak to get started.

Fill in your answer here

Maximum marks: 5

9 AE vs VAE

What are autoencoders (AE) and variational autoencoders (VAE)? How do they differ? How does one choose which one is more appropriate for the task at hand?

Fill in your answer here

Maximum marks: 5

10 Back-propagation

Consider the following regression task. We want to predict salaries of current students in their first job after graduation (y_1) and 10 years after graduation (y_2). We have collected 10 features about the studies and followed-up former students to get salary information. In other words, we have a 10-dimensional feature vector $x \in \mathbb{R}^{10\times 1}$
and a two-dimensional label $y = (y_1, y_2)$.

We have the following neural network with one hidden layer and **two** output layers:

$$z^{[1]} = W_0 x + b_0$$
$$a^{[1]} = relu(z^{[1]})$$
$$\hat{y}_1 = W_1 a^{[1]} + b_1$$
$$\hat{y}_2 = W_2 a^{[1]} + b_2$$

where $W_0 \in \mathbb{R}^{100\times 10}$, $b_0 \in \mathbb{R}^{100\times 1}$, $W_1 \in \mathbb{R}^{1\times 100}$, $b_1 \in \mathbb{R}$, $W_2 \in \mathbb{R}^{1\times 100}$and $b_2 \in \mathbb{R}$ are the parameters of the network.

As the loss function, we use MSE for both output layers. That is,

$$L(y, \hat{y}) = \alpha(\hat{y}_1 - y_1)^2 + (1-\alpha)(\hat{y}_2 - y_2)^2$$

where $\alpha \in [0, 1]$ is a hyperparameter that determines the relative importance of the objectives.

Tasks:

1. Draw the computational graph.
2. Compute the gradient of the loss function with respect to the parameters using back - propagation. You can assume that you have only one data point.
3. Instead of creating this complicated network, one could learn two separate networks, one for each task. Why on earth would anybody want to combine these two tasks instead of learning two separate networks?

Derivation rules:

$\mathbf{f(x)}$	$\mathbf{f'(x)}$
x^d	dx^{d-1}
$relu(x)$	$\begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \\ undefined & \text{if } x = 0 \end{cases}$

$$(f+g)' = f' + g'$$
$$(fg)' = f'g + fg'$$
$$f(g(x))' = f'(g(x))g'(x)$$

Hint: Use a shorthand for the derivative of ReLU (because we do not know whether the input is positive or negative).

Hint: Some parts of the gradient are derived in an identical way as some other parts. In such cases, it is enough to derive the gradient for one part and observe (and state) that the other part is done in similar fashion.

It is recommended that you solve this question on paper.

Fill in your answer here

Format

Words: 0

Maximum marks: 10

11 Obligatory projects

This is a placeholder for points from the projects.

Maximum marks: 45