General information about the exam

This exam has 8 questions (You can see 9 questions in Inspera but Question 9 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.

No examination support material is allowed on this exam.

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 8.

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 examination.

In the 15 minutes after exam end time, you can fill out other requested information at the top of the page: date, your candidate number, course code, number of pages etc. 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: $f_A(z) = 8z - 3$
Bertha: $f_B(z) = \min(0, -z)$
Cecilie: $f_C(z) = \begin{cases} 0 & \text{if } z \leq 0 \\ 1 & \text{otherwise} \end{cases}$

What are pros and cons of each suggestion?

Fill in your answer here

Format

Words: 0

Maximum marks: 5

2 What could possibly go wrong?

Thanks to your excellent performance at the INF265 exam in spring 2023, you have ended up as a group leader at the same course in spring 2024. As a group leader, your task is to help students at group sessions. Lise is not sure whether she is doing everything correctly and she comes to a group session for help.

Read Lise's explanation below and give her advice.

Lise:

I have implemented my own gradient descent algorithm.

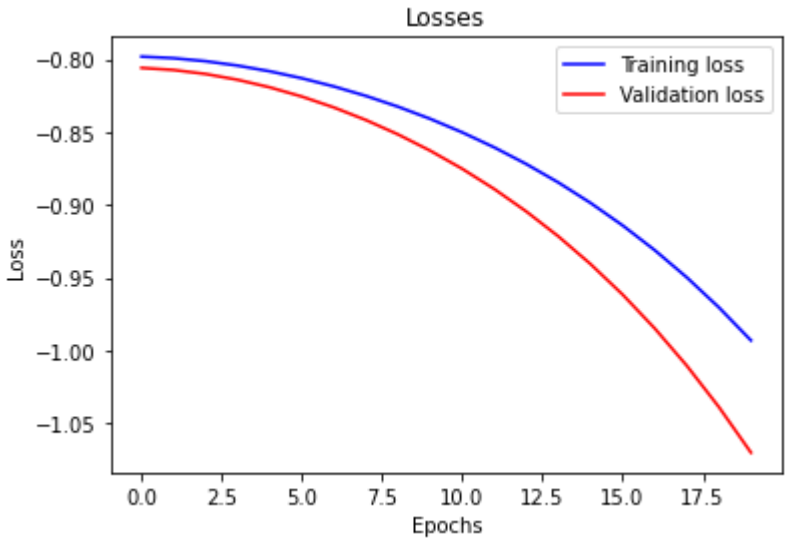
I am learning a simple neural network solving a binary classification problem. I have data with 2 features. 40% of the data points belongs to class 1 and 60% to class 0. My training set consists of 500 data points and validation set 200 points.

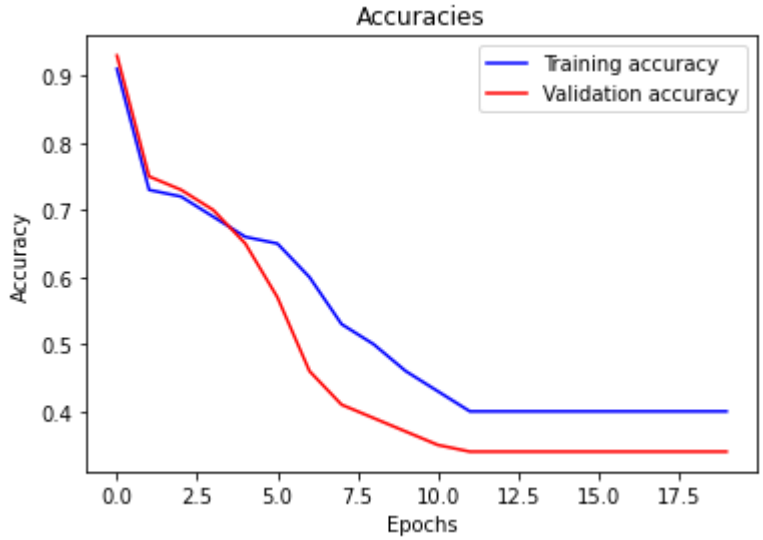
The network has two hidden layers with 8 neurons each. I use ReLU as an activation function. The output layer has a sigmoid function.

I implemented the cross-entropy loss using the following formulas:
$L(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^{n} l(\mathbf{y}^{(i)}, \hat{\mathbf{y}}^{(i)})$
and
$l(y, \hat{y}) = \sum_{j=1}^{q} y_j \log \hat{y}_j.$
The latter formula is loss for one data point. There are q different class labels. y is a one-hot vector representing the true label ($y_j = 1$ only if the true label is the jth class). $\hat{y}$ is the prediction where $\hat{y}_j$ is the probability of the jth class. The total loss (the first formula) is just a sum of losses for each data point (n data points).

I used learning rate 0.5, momentum with parameter 0.9 and weight decay with parameter 1. I trained my model for 20 epochs. I had all training points in one batch.

I have plotted both losses and accuracies for both training and validation sets; see below.





Is everything working as it should?

Fill in your answer here

Format

Words: 0

Maximum marks: 5

3 Regularisation

Answer briefly to the following questions:

1. What is regularisation?
2. Why do we use regularisation?
3. Give examples of regularisation techniques used with neural networks.

Fill in your answer here

Format

Words: 0

Maximum marks: 6

4 GANs

Give short answers to the following questions:

1. What are generative adversarial networks (GANs) used for?
2. Describe briefly the main parts of a GAN model.
3. Describe briefly how one trains a GAN.

Fill in your answer here

Format

Words: 0

Maximum marks: 5

5 Attention

Answer briefly to the following questions:

1. What is the attention mechanism and how does it work?
2. What are the main benefits of the attention mechanism compared to recurrent neural networks?

Fill in your answer here

Format

Words: 0

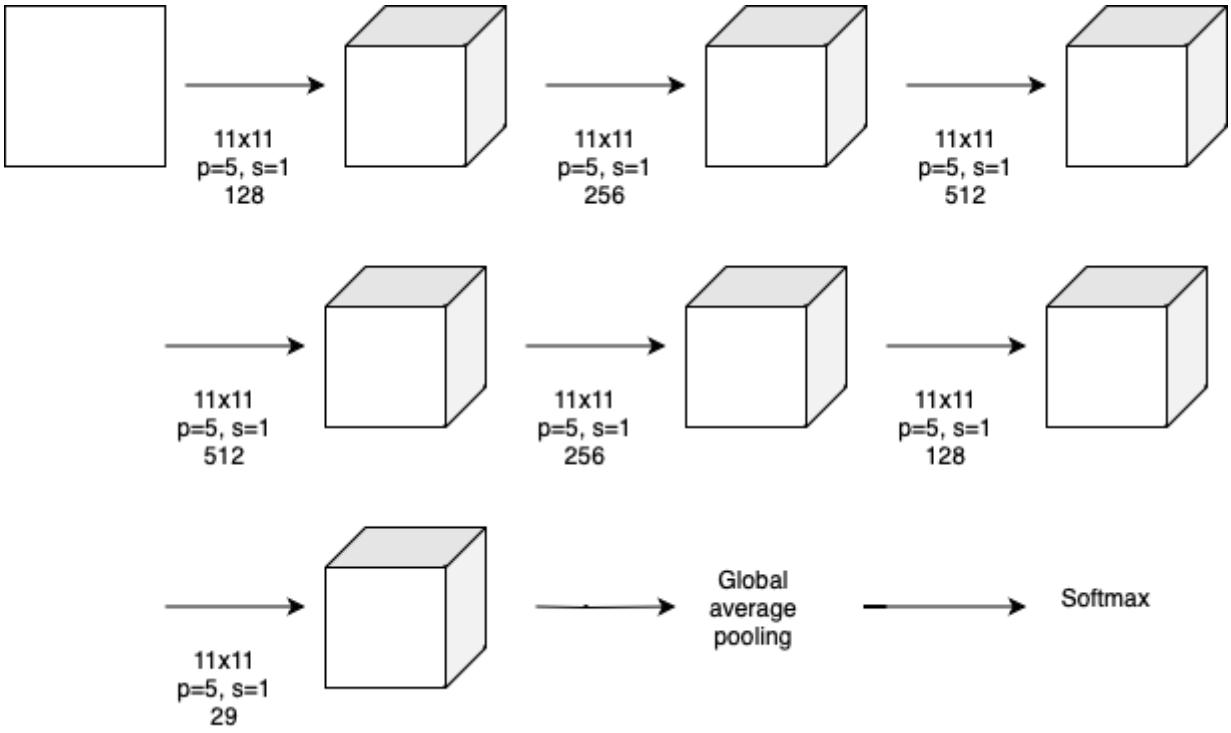
Maximum marks: 8

6 Convolutional network

Ingvild has got a new state-of-the-art GPU and she is excited because it allows her to train large neural networks.

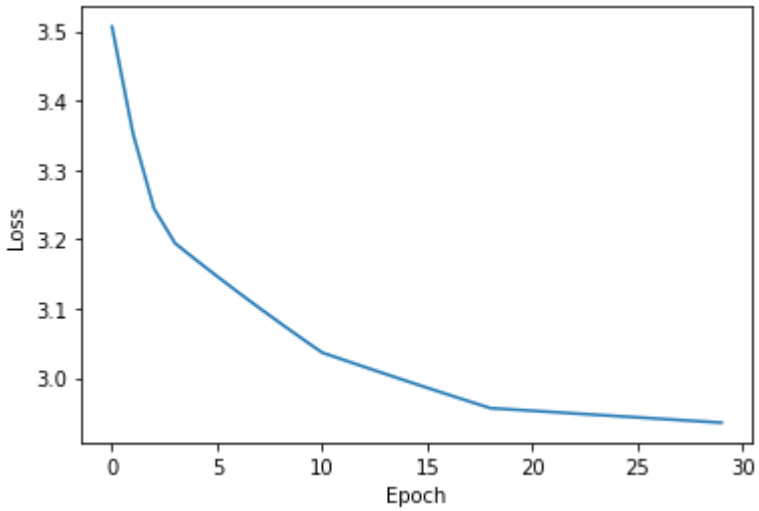
She starts by building a model for classifying handwritten letters. As an input, the model takes a black-and-white image of a handwritten letter in Norwegian alphabet (one input channel). Most of the images are 64x64 pixels but some are smaller and some larger. The output of the model is a class label A, ..., Å (There are in total 29 classes). Ingvild has collected large data set. She has 10,000 labelled data points for each class totalling 290,000 data points.

The architecture of Ingvild's network can be seen below. It consists of seven convolutional layers and average pooling and softmax layers in the end. For each convolutional layer, the figure shows the size of the filters (first row), padding and stride (second row) and the number of filters (third row). All convolutional layers have sigmoid activation functions.



Ingvild wants the model to be able to make predictions for varying input sizes and thus she uses global average pooling. Global average pooling takes an average over all pixels in the activation map and outputs one number per channel.

Ingvild trains the model with batch size 1024 and learning rate 0.000001. She plots cross-entropy loss on the training data after each epoch; see below.



Despite of her powerful GPU, the model performs quite poorly on the training set.

Help Ingvild. What are potential sources of the disappointing performance? What could she try in order to improve performance?

Fill in your answer here

Format

Words: 0

Maximum marks: 8

7 Modelling help

The department wants to evaluate how study success relates to job placement. Specifically, they are interested in students taking machine learning courses.

The department has collected a data set from students who have taken INF265 in the past. Some variables are about past studies of the student (Number of informatics courses taken prior the course, number of mathematics courses taken prior the course, ...) and some are about what the student did during INF265 (Number of homework returned, points from the exam, ...).

The points that students have got from projects at INF265 are somehow lost but free-text fields written by graders are available. These fields contain short descriptions about the overall quality of the projects. Examples of comments: "excellent", "ok", "good", ...

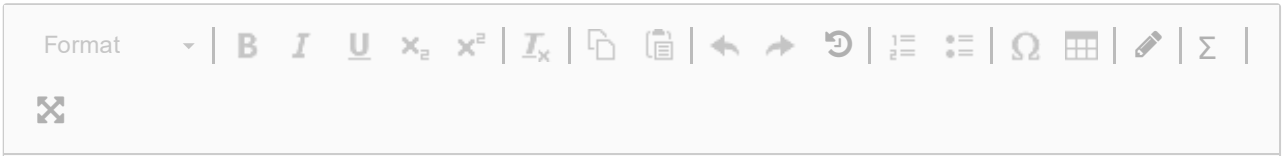
The department wants to predict two targets. The targets are the type of work that the student is doing five years after graduation (data scientist, software developer, consultant, ...) and salary five years after graduation.

Currently, the department uses a random forest model that works well. However, they would like to know whether a neural network would do an even better job.

Help the department. **Answer the following questions**:

1. Give the department advice on how to develop a neural network model for the task. How to preprocess data? Suggest an architecture. Define a loss function.
2. How do we know whether the neural network is better than the random forest? How would you evaluate the models? What performance measure would you use?

Fill in your answer here



Words: 0

Maximum marks: 8

8 Forward and backward propagation

Consider the following neural network with one hidden layer for a regression task. The network has weight matrices $W^{(1)} = \begin{bmatrix} 1 & 3 & -1 \\ 2 & 1 & 1 \end{bmatrix}$, $W^{(2)} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$, $b^{(1)} = [0 \quad 1 \quad -3]$ and $b^{(2)} = 0$ where $W^{(1)}$ and $b^{(1)}$ are the parameters of the hidden layer and $W^{(2)}$and $b^{(2)}$are the parameters of the output layer. The hidden layer has a ReLU activation function.

Answer the following questions:

A) Forward pass. Suppose we perform a forward pass with the input $x = [1 \quad -1]$. What is the output of the network? Show intermediate results.

B) Loss. Consider the squared loss $L = \frac{1}{2}(\hat{y} - y)^2$ where $\hat{y}$ is the prediction of the network. Suppose the true label is $y = 5$. What is the loss?

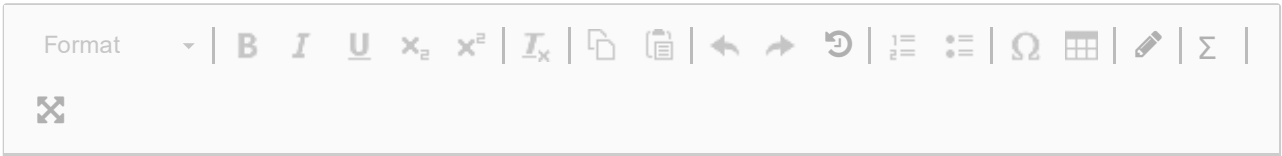
C) Backward pass. Compute the partial derivatives of the parameters using back-propagation. Show your work.

D) Gradient descent update. Update the parameters using gradient descent with learning rate 0.1.

E) Here we did a gradient descent update using a mini-batch with size 1. Why do we usually prefer larger mini-batches?

It is recommended to solve this task on paper. See instructions under "General information".

Fill in your answer here



Words: 0

Maximum marks: 10

9 Points from projects

This is a placeholder for points from the projects.

Maximum marks: 45