UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	IN 54000/IN 9400 — Machine Learning for Image Analysis
Day of examination:	June 3 2020
Examination hours:	9.00 June 3 – 9.00 June 10

This exercise set consists of 11 pages.

Appendices:	None
Permitted aids:	All

Read the entire exercise text before you start solving the exercises. Please check that the exam paper is complete. If you lack information in the exam text or think that some information is missing, you may make your own assumptions, as long as they are not contradictory to the "spirit" of the exercise. In such a case, you should make it clear what assumptions you have made.

Read the following link for important exam information: https://www.mn.uio.no/english/about/hse/corona/examination-2020.html

You will submit a single PDF file. In this file, the exercises should be answered in given order. Any figures or drawings MUST be included at the correct place in the document. If you make sketches or computations on paper, include photo of them in good quality and with good resolution, and place the photo under the appropriate subtask.
If you use any source different from the course material to answer a question, include full reference to the source in the answer of the corresponding subtask. Use proper styles for citing wepages or papers.
Most of your answers should include a discussion, typically a few sentences.

Every subtask has equal weight in the evaluation.

(Continued on page 2.)

About these solution hints

This year's exam was a home exam, and the questions are adapted to that to allow individual formulations and reflections. Some of the questions are a bit open, and valid arguments not listed in the solutions hints were also given full score during the exam grading.

(Continued on page 3.)

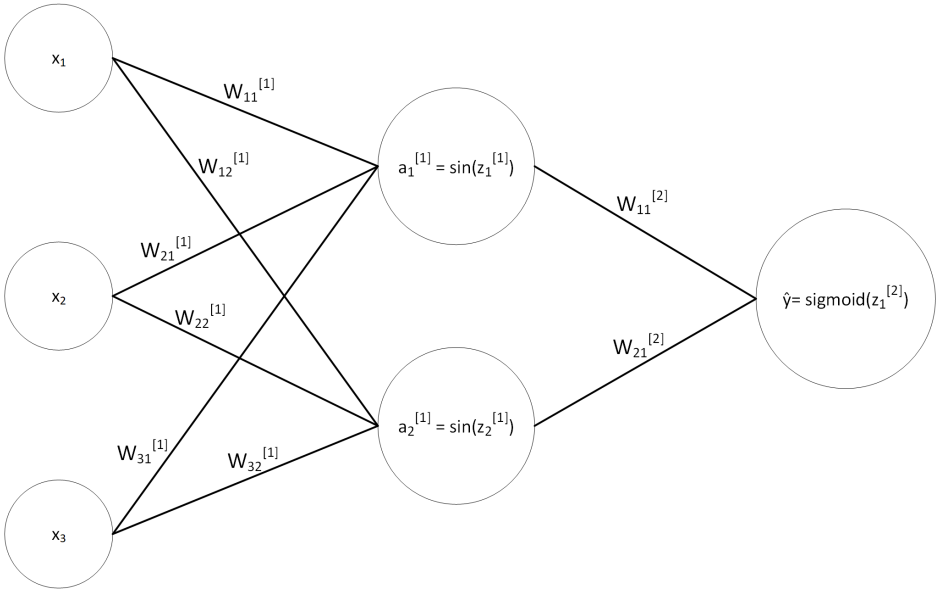
Exercise 1 Computations for a simple network

You are given a simple feedforward network with one fully connected hidden layer with 2 nodes, activation function sin(z) (assume z is measured in radians), and a binary classification problem. The output layer uses sigmoid activation. The input vector to the network is $x = [x_1, x_2, x_3]$.

1a

Make a sketch of a computational graph for the network (from the inputs to the output $\hat{y}$.

Note that going from x to z or a can be drawn using one or two operator cell, as long as the content of each is clearly written



1b

Let $x_1 = 2, x_2 = 1, x_3 = 3$
We assume that the bias terms are zero, and the initial values of the weights are:
$W_{11}^{[1]} = 1, W_{12}^{[1]} = 2, W_{21}^{[1]} = 1, W_{22}^{[1]} = 3, W_{31}^{[1]} = 2, W_{32}^{[1]} = 1,$
$W_{11}^{[2]} = 3, W_{12}^{[2]} = 4.$
Compute the predicted value $\hat{y}$. Show your computations.

Answer: Let g(z) denote the sigmoid function

$$z_1^{[1]} = 2 * 1 + 1 * 1 + 3 * 2 = 9, z_2^{[1]} = 2 * 2 + 1 * 3 + 3 * 1 = 10$$

$$\hat{y} = g(3 * \sin(z_1^{[1]}) + 4 * \sin(z_2^{[1]})) = g(-0.94) = 0.2810$$

Note that sin(z) is calculated assuming that z is in radians.

(Continued on page 4.)

1c

If we use a logistic cost function, and the true output for the single sample is $y = 1$, compute the first update of $W_{11}^{[2]}$ if we use gradient descent with a learning rate of 0.2.

Include all your computations.

The logistic loss function is

$$l(w,b) = -\sum_{i=1}^{m} y^{(i)} \log \hat{y} + (1 - y^{(i)}) \log(1 - \hat{y})$$

Here, we have a single sample with y=1 and this reduces to:
$l(w,b) = -\log(\hat{y})$

$$\frac{\partial l}{\partial W_{11}^{[2]}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_1^{[2]}} \frac{\partial z_1^{[2]}}{\partial W_{11}^{[2]}} \tag{1}$$

$$\frac{\partial l}{\partial \hat{y}} = -\frac{1}{\hat{y}} \tag{2}$$

$$\frac{\partial \hat{y}}{\partial z_1^{[2]}} = \hat{y}(1 - \hat{y}) \tag{3}$$

$$\frac{\partial z_1^{[2]}}{\partial W_{11}^{[2]}} = a_1^{[1]} \tag{4}$$

$$\frac{\partial l}{\partial W_{11}^{[2]}} = -(1 - \hat{y}) a_1^{[1]} = -(1 - 0.2810)\ \sin(9) = -0.296 \tag{5}$$

With a learning rate of 0.2 we get the update :
$W_{11}^{[2]} = W_{11}^{[2]} - 0.2 * (-0.296) = 3.0592$

1d

Discuss if sin(z) would be a good activation function. Give your views on some properties that an activation function should have.

Here we are looking for arguments. No strict correct answer exist.

- Pros:
  - The sin function is bounded between 0 and 1.
  - Differentiable for all values
  - Zero centered
- Cons:

(Continued on page 5.)

- It can be problematic that the function is periodic. However, this could be handled by normalizing the data to lie within half a period (e.g $-\pi/2$ to $\pi/2$).
- More computational expensive than the relu activation function.

1e

Explain why the range or standard deviation of the weights in a layer in a neural network should depend on the size of the input to the layer.

Each output is a weighted sum of the number of inputs. If the size of the weights does not depend on the size of the input, the scale of the output would depend on the number of inputs. Many inputs would mean growing output values, and few would mean vanishing outputs In the end, this is to avoid exploding/vanishing gradients, and control convergence speed.

Exercise 2 Convolutional neural networks

2a

Explain briefly why fully connected neural networks do not work well for image classification.

The two main arguments are:

- A fully connected neural network is not translational invariant.
- A fully connected neural network is also parameter inefficient as it does not reuse weights.

2b

Argue whether the effective receptive field (field of view) is larger or smaller than the theoretical receptive field.

We can count the number of paths (connections) from the input data to a neuron. Input data located right below a neuron have more paths compared to input data located at the edge of the receptive field. The number of paths can be viewed as a weighting function, hence the effective receptive field will be smaller than the theoretical receptive field.

(Continued on page 6.)

2c

Discuss whether a CNN is efficient for detecting long range dependencies in an image.

We typically need many CNN layers to be able to connect distant information within an image. We can use pooling/stride and dilation to more efficiently increase the receptive field, but in general CNN's are inefficient for detecting long range spatial dependencies.

2d

Consider the model architecture consisting of a CNN followed by two dense (fully-connected) layers. Discuss whether the model is invariant to translation.

The spatial location of the features (within a feature map $C \times N_x \times N_y$) produced by a CNN depends on the spatial location of the objects in the input image. Flattening the CNN feature map (to $C \times N_x \cdot N_y$) and passing the features to a fully connected layer breaks the translational invariance. However, if global average pooling would be used instead of the two fully connected layers the model would be translational invariant.

Exercise 3 Recurrent neural networks

3a

How could you structure a CNN to be used for 1D time series data with a stream of incoming data?

Note that the causality of the data is important. Only data seen at a given time t or previous timepoints can be used. We can use temporal convolutions (causal convolutions). The output is computed using data from the past only. If a small delay is acceptable in the application, buffering is also possible.

3b

Discuss when you would consider to use a bi-directional RNN.

A bidirectional RNN can be helpful when the prediction is dependent on information about the "past" and the "future" (both directions).

(Continued on page 7.)

Exercise 4 Statistical properties of classification

4a

Discuss briefly challenges associated with a binary classification problem where you have labelled data from 9900 samples for class 1 and 100 samples for class 2.

There are several challenges, both to get the network to learn both classes, and when computing the classification accuracy.

There are also other valid arguments, e.g. related to dead neurons with ReLU.

If the first N samples are all the majority class, and the learing rate is high, it might push ReLU to a point from which the network cannot recover.

4b

Explain what a minibatch is, and discuss any potential drawbacks of using either a very small or a very large minibatch size.

A minibatch is the collection of samples used to compute gradients when updating using SGD during backpropagation. The major drawbacks are:
Small minibatch: too small for proper batchnorm, speed (underutilized GPU), convergence issues.

Large minibatch: memory constraints. Little stochasticity in SGD could be argued.

4c

Discuss what to expect if the training and test data is drawn from different sample distributions (processes).

You should expect the model to not generalize well, or at least, you have no control on the generalization error.

4d

What can be a drawback of extreme hyper-parameter tuning?

As explained in the comments of questions during the exam, by extreme hyper-parameter tuning we mean tuning too much (or too little, this could also be a valid interpretation of the question). Extreme hyper-parameter tuning can result in over-fitting on the validation data set and lead to poor generalization.

(Continued on page 8.)

4e

PhD students only

Discuss L2 regularization vs. multitask learning as regularization approaches.

Regularization limits the number of potential network weight configurations. The reduction in likely hypotheses can thus reduce the generalization error. L2 regularization has the effect of lowering the values of the network weights and making the decision boundary smoother. Forcing the network weights to be small may not always be beneficial. Training the model simultaneously on related tasks can infer more general features which hence reduce the generalization error.

Exercise 5 Object detection and segmentation

Assume that you start out with an already-trained convolutional network (CNN) for image classification. You are then given the task of detecting and localizing objects of interest in images. You are interested in the following three object types: tigers, leopards, and lions.

5a

Initially, assume that there is at most one such object of interest in each image. Explain how you would modify the network to detect and localize this object.

Keywords: eight additional outputs, the 3-classes plus background, and 4 numbers for the bounding box. Loss function: add a L2-loss for the bounding box in addition to the cross entropy loss.

5b

Let us now consider cases where you might have multiple objects of interest within the same image. One such appoach is the sliding window approach. Discuss some challenges with this approach.

Keywords: computational requirements, the difficulty of getting enough spatial context

5c

Another approach to handle multiple objects within each image is to let your network produce multiple/many bounding boxes and class labels. Discuss challenges with this approach, and how you can solve this.

Challenge: must avoid that not all detect and predict the same object. A solution is to use anchor boxes, which gives a prior or default box.

(Continued on page 9.)

Exercise 6 Selected topics

6a

Describe an example from the course on how we can use up-sampling in convolutional networks.

One example: Unet, but we have also mentioned it for visualization

6b

Describe briefly a method for visualizing or interpreting what a trained network has learned.

Several possible methods like Guided Backprop, saliency maps, CAM/GradCam, Layerwise relevance propagation. A small description of the choosen method is expected.

6c

Describe briefly the difference between an autoencoder and a variational autoencoder for the problem of data generation.

Autoencoders do not extrapolate beyond the distributions seen in the training data. Varational autoencoders partly fills the space around samples by only estimating a distribution in latent space, and adding noise (in latent space).

6d

Find an application of deep learning that in your opinion raises some ethical aspects. Describe the application and discuss the ethical concerns.

Many choides are possible, and concerns can be application, bias in the data, bias the algorithm, legal issues, privacy et. al. The main point is reflections from the candidate.

6e

PhD students only Find a paper that compares ReLU and LeakyReLU. Describe briefly their findings, the data set, and the difference in performance the paper reports.

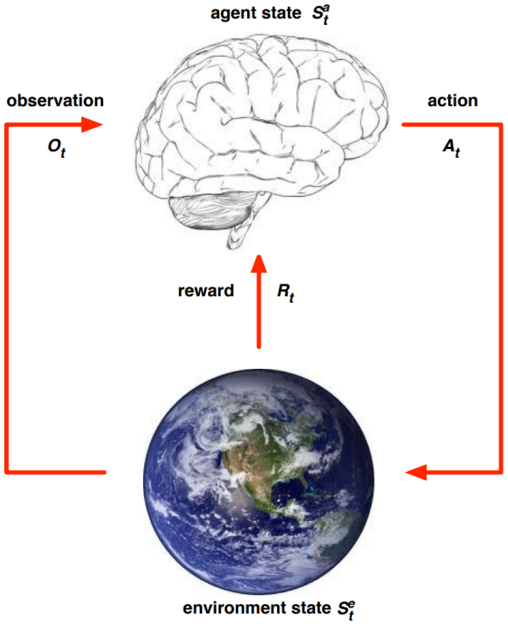
There are several papers. All answers giving a citation and summarizing the main findings were given full score.

Exercise 7 Reinforcement learning- 2020

7a

Explain the quantities, S_t^a, A_t, S_t^e, R_t, and O_t in the figure:

(Continued on page 10.)



- S_t^a - is the agents representation/state/understanding of the world (system). The subscript, t, indicate at time step t. The agent's state can be e.g. a table, a neural network, or the environment state in a full observatory system. The agent selects an action based on the agent state.
- A_t - is the action selected by the agent at time step t.
- S_t^e - is the environment state. The environment state is the true configuration of the world (system).
- R_t - is the reward sampled by the environment state given.
- O_t - is the observation done by the agent of the environment state. The observation can include full knowledge of environment state, but can also be a practical observation (sensor data).

7b

Explain the trade-off between exploration and exploitation.

Reinforcement learning is different compared to supervised learning as the agent affects the training data itself. The agent select actions and based on these actions the training data is generated.

In the Bellman (optimality) equation the q-value is defined by selecting the future actions with the highest q-value. For an untrained agent, the action with the highest q-value can be incorrect and result in the agent only visiting part of the state-space. For this reason, early in the training

(Continued on page 11.)

phase (of the agent) it is common to "explore" with selecting a random action with probability ϵ.

For a trained agent it is inefficient to keep exploring sub-optimal regions of the state-space. We can exploit the knowledge of the agent to search in already promising regions of the state-space by selecting promising actions.